5-2009

Testing and Estimation for Functional Data with Applications to Magnetometer Records

Inga Maslova
Utah State University

TESTING AND ESTIMATION FOR FUNCTIONAL DATA WITH APPLICATIONS TO MAGNETOMETER RECORDS

by

Inga Maslova

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences

Approved:

Dr. Piotr Kokoszka Major Professor	Dr. Daniel Coster Committee Member
Dr. Richard Cutler Committee Member	Dr. Peg Howland Committee Member
Dr. Lie Zhu Committee Member	Dr. Byron R. Burnham Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2009

ABSTRACT

Testing and Estimation for Functional Data with Applications to Magnetometer Records

by

Inga Maslova, Doctor of Philosophy

Utah State University, 2009

Major Professor: Dr. Piotr Kokoszka
Department: Mathematics and Statistics

The functional linear model, $Y_n = \Psi X_n + \varepsilon_n$, with functional response and explanatory variables is considered. A simple test of the nullity of Ψ based on the principal component decomposition is proposed. The test statistic has asymptotic chi-squared distribution, which is also an excellent approximation in finite samples. The methodology is applied to data from terrestrial magnetic observatories.

In recent years, the interaction of the auroral substorms with the equatorial and mid-latitude currents has been the subject of extensive research. We introduce a new statistical technique that allows us to test at a specified significance level whether such a dependence exists, and how long it persists. This quantitative statistical technique, relying on the concepts and tools of functional data analysis, uses directly magnetometer records in one minute resolution, and it can be applied to similar geophysical data which can be represented as daily curves. It is conceptually similar to testing the nullity of the slope in the straight line regression, but both the regressors and the responses are curves rather than points. When the regressors are daily

high latitude H–component curves during substorm days and the responses are daily mid– or low latitude H–component curves, our test shows significant dependence (the nullity hypothesis is rejected), which exists not only on the same UT day, but also extends into the next day for strong substorms.

We propose a novel approach based on wavelet and functional principal component analysis to produce a cleaner index of the intensity of the symmetric ring current. We use functional canonical correlations to show that the new approach more effectively extracts symmetric global features. The main result of our work is the construction of a new index, which is an improved version of the existing wavelet-based index (WISA) and the old Dst index, in which a constant daily variation is removed. Here, we address the fact that the daily component varies from day to day and construct a "cleaner" index by removing non-constant daily variations.

A wavelet-based method of deconvoluting the solar quiet variation from the low and mid-latitude H-component records is proposed. The resulting daily variation is non–constant, and its day–to–day variability is quantified by functional principal component scores. The procedure removes the signature of an enhanced ring current by comparing the scores at different stations. The method is fully algorithmic and is implemented in the statistical software R.

R package for space physics applications is developed. It consists of several functions that compute indices of the storm activity and estimate the daily variation. Storm indices are computed automatically without any human intervention using the most recent magnetometer data available. Functional principal component analysis techniques are used to extract day-to-day variations. This package will be publicly available at Comprehensive R Archive Network (CRAN).

(154 pages)

I dedicate this thesis to my dear mom.

ACKNOWLEDGMENTS

I would like to thank my adviser, Piotr Kokoszka, for the guidance over the past six years, for teaching me independence and creativity in the research process.

I would also like to thank my committee members, Daniel Coster, Richard Cutler, Peg Howland, Lie Zhu, as well as Jan J. Sojka, for valuable interactions during my program.

Many thanks to Andrejus Parfionovas for all the technical support and valuable insights he has provided; and to Robertas Gabrys for useful discussions and encouragement.

I am grateful to my husband, Andres Ticlavilca, for making this time a little less stressful.

I would like to thank my parents and my sister for supporting and inspiring me.

Research was partially supported by NSF grants DMS-0413653 and DMS-0804165. Data was provided by USGS via the global network of observatories INTERMAGNET.

Inga Maslova

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Functional data analysis (FDA) is an area of statistics that has been around for about twenty years. Historically, functional data were analyzed using time series and multivariate methods at discrete points. However, most of the data available are continuous and it is more natural to analyze it as functional data. Functional methods can be applied even for the data recorded at irregular intervals, which is an important advantage of this approach (e.g. [1]). In addition to that functional data can be evaluated at any point, which allows to use methods requiring evenly-spaced observations.

Functional data can come in many forms and arise in many different fields. The main defining quality of such data is that they consist of functions – curves. Examples include meteorological and pollution monitoring data, seismic data, economic data collected either over many years or minute records of stock prices, etc.

This work was motivated by magnetometer records. We analyze the intensity of the Earth's magnetic field data that comes from the ground-based magnetometers. A magnetometer is a device that measures three components of the magnetic field at a fixed location. In this work the horizontal (H) component is analyzed ([2]). There are over a hundred ground-based magnetic stations, most of which are equipped with digital magnetometers. These magnetometers record the strength and direction of the field every five seconds. However, the magnetic field exists at any moment of time, so it is natural to think of a magnetogram as an approximation to a continuous record. The raw magnetometer data are cleaned and reported as averages over one minute intervals.

Statistics is concerned with obtaining information from a sample of observations $X_1, X_2, \ldots, X_N$. The X_n can be scalars, vectors or other objects. For example, each X_n can be a satellite image, in some spectral bandwidth, of a particular region of the Earth taken at time n. Functional Data analysis (FDA) is concerned with observations which are viewed as functions defined over some set T. A satellite image processed to show surface temperature can be viewed as a function X defined on a subset T of a sphere, $X(t)$ being the temperature at location t. The value $X_n(t)$ is then the temperature at location t at time n. Clearly, due to finite resolution, the values of X_n are available only at a finite grid of points, but the temperature does exist at every location, so it is natural to view X_n as a function defined over the whole set T.

The data that motivated the research presented in this dissertation is of the form $X_n(t),\ t \in [a, b]$, where $[a, b]$ is an interval on the line. Each observation is thus a curve. These curves can arise in many ways. Figure 1.1 shows a reading of a magnetometer over a period of one week. Panel (a) of the figure shows the magnetic field intensity records at one minute resolution. The dotted vertical lines separate days in Universal Time (UT). It is natural to view a curve defined over one UT day as a single observation because one of the main sources influencing the shape of the record is the daily rotation of the Earth. When an observatory faces the Sun, it records the magnetic field generated by wind currents flowing in the ionosphere which are driven mostly by solar heating. Thus, panel (b) of Figure 1.1 shows seven consecutive functional observations.

Many important examples of data that can be naturally treated as functional come from financial records. Figure 1.2 shows two consecutive weeks of Microsoft stock prices in one minute resolution. In contrast to the magnetic field, the price of an asset exists only when the asset is traded. A great deal of financial research

has been done using the closing daily price, i.e. the price in the last transaction of a trading day. However many assets are traded so frequently that one can practically think of a price curve that is defined at any moment of time. The Microsoft stock is traded several hundred times per minute. The values used to draw the graph in Figure 1.2 are the closing prices in one-minute intervals.

It is natural to choose one trading day as the underlying time interval. If we do so, Figure 1.2 shows 10 consecutive functional observations. From these functional observations, various statistics can be computed. For example, the top panels of Figure 1.3 show the mean functions for the two weeks computed as $\hat{\mu}(t) = 5^{-1} \sum_{i=1}^{5} X_i(t)$, where $X_i(t)$ is the price at time t on the ith day of the week. We see that the mean functions have roughly the same shape (even though they have different ranges), and we may ask if it is reasonable to assume that after adjusting for the ranges, the differences in these curves can be explained by chance, or these curves are really different. This is clearly a setting for a statistical hypothesis test which requires the usual steps of model building and inference. The bottom panels of Figure 1.3 show the five curves $X_i(t) - \hat{\mu}(t)$ for each week. We will often work with functional data centered in this way, and will exhibit the curves using the graphs as those in the bottom panels of Figure 1.3.

Functional data arise not only from finely spaced measurements. For example, when measurements on human subjects are made, it is often difficult to ensure that they are made at the same time in the life of a subject, and there may be different numbers of measurements for different subjects. A typical example are growth curves, i.e. $X_n(t)$ is the height of subject n at time t after birth. Even though every individual has a height at any time t, it is measured only relatively rarely. Thus it has been necessary to develop methods of estimating growth curves from such sparse unequally spaced data, in which smoothing and regularization play a crucial role. Examples and

methodology of this type are discussed in the monographs of [3] and [4]).

It is often useful to treat as functional data measurements that are neither sparse nor dense. Figure 1.4 shows the concentration of nitrogen oxide pollutants, referred to as NO_x, measured at Barcelona's neighborhood of Poblenou. The NO_x concentration is measured every hour, so we have only 24 measurements per day. It is nevertheless informative to treat these data as a collection of daily curves because the pattern of pollution becomes immediately apparent. The pollution peaks in morning hours, declines in the afternoon, and then increases again in the evening. This pattern is easy to explain because the monitoring station is in a city center, and road traffic is a major source of NO_x pollution. Broadly speaking, for functional data the information contained in the *shape* of the curves matters a great deal.

This dissertation consists of five parts. Each of them is an individual manuscript that deals with functional data analysis techniques and wavelet analysis for time series. Chapter 2 deals with the development of a test for no effect in the fully functional linear model. In Chapter 3, the latter test is applied to magnetometer records and reveals some unexpected findings. Chapters 4 and 5 introduce a novel methodology applying wavelet and functional data analysis techniques to space physics data. Finally, in Chapter 6 the description of the R package is provided. A brief description of these chapter now follows.

Chapter 2 introduced a test procedure for the fully functional linear model. The functional linear model is probably one of the most popular models of FDA analysis. It is described in its various forms in Chapters 12–17 of [4], a general overview is given in [5]. To name a few references of both theoretical and applied flavors one should mention [6], [7], [8], [9], [10], [11], [12], [13], [14].

This model is defined by the equation

$$Y_n = \Psi X_n + \varepsilon_n, \quad n = 1, 2, \ldots, N.$$

Testing the lack of dependence hypothesis is equivalent to testing in Ψ is zero. The test procedure proposed in Chapter 2 is similar to that introduced in [15]. However, the data that motivated this research consists of pairs of curves, therefore, it puts us into fully functional linear model context.

The work given in Chapter 3 has been motivated by this conjecture and shows that this is indeed the case by using a statistical test of significance proposed by [16]. It is believed, [17], that the auroral currents may have an indirect impact on the equatorial and mid-latitude currents. We test if observations at mid- and low-latitude stations are independent of the substorms that are recorded at high-latitude stations.

In Chapter 4 a novel procedure of removing a nonconstant daily variation from the ground based magnetometer data is introduced. This work was motivated by ideas introduced in [18] who use the method of natural orthogonal components to analyze the daily magnetic variation. We also build on the work of [19] who developed the automated procedure of extracting the ring current index (WISA). Yet, in WISA and other indexes, like traditional Dst, the constant daily variation is removed. The approach we propose involves wavelet and functional principal component methods. The nonconstant daily variation is removed and a cleaner storm index is constructed.

Chapter 5 deals with estimating the Solar quiet component from the H-component of the magnetometer records. Here, we introduce a procedure which is an improved version of the algorithm introduced in [20].

Chapter 6 provides a brief overview of the R package, which includes the techniques introduced in this work.

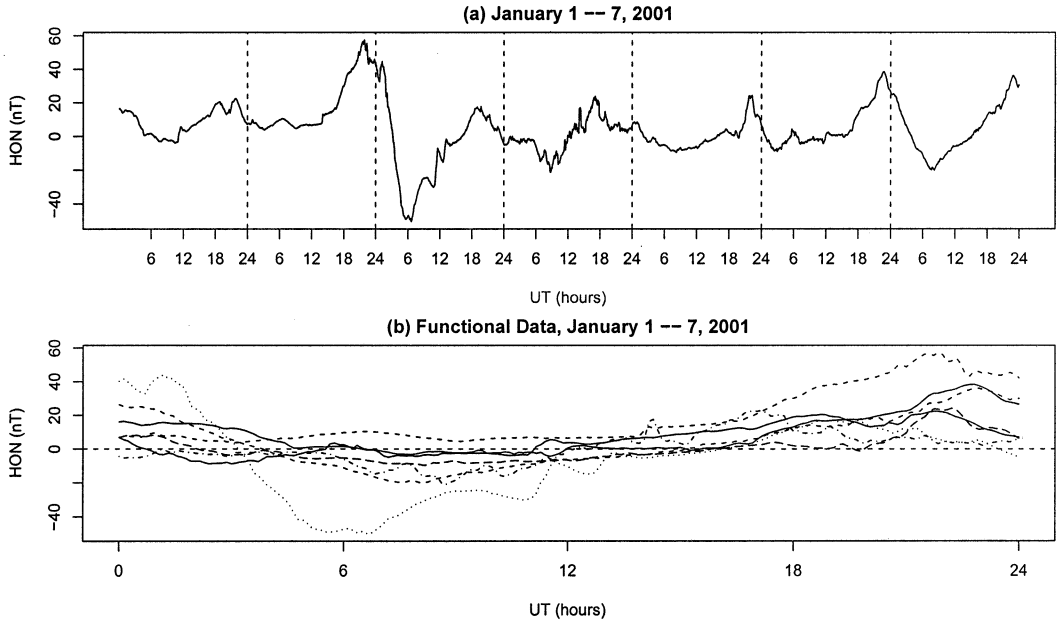


Fig. 1.1: (a) The horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT, (b) Same observations as in panel (a) presented as functional daily magnetic field activity

Fig. 1.2: Microsoft stock prices in one-minute resolution, May 1-5, 8-12, 2006

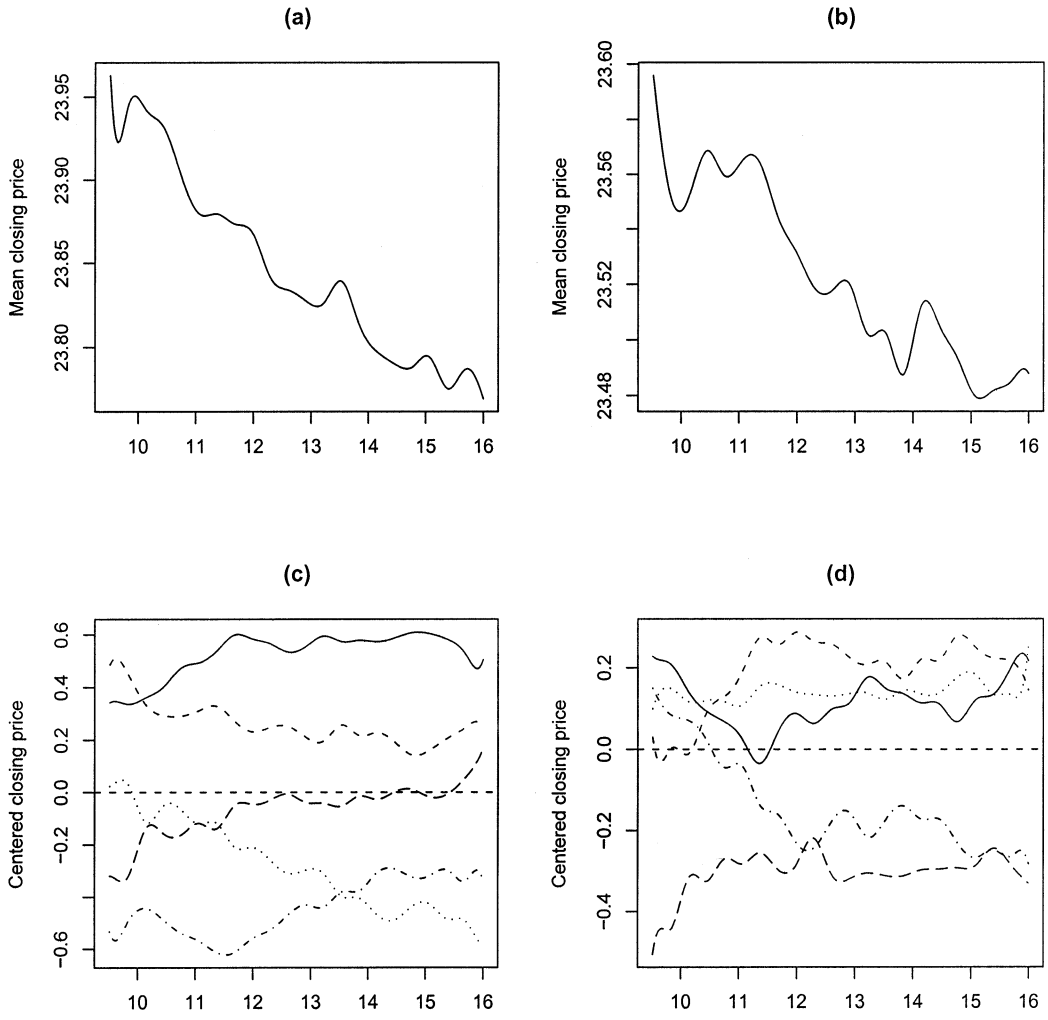


Fig. 1.3: (a) Mean function of Microsoft stock prices, May 1-5, 2006; (b) Mean function of Microsoft stock prices, May 8-12, 2006; (c) Centered prices of Microsoft stock, May 1-5, 2006; (d) Centered prices of Microsoft stock, May 8-12, 2006

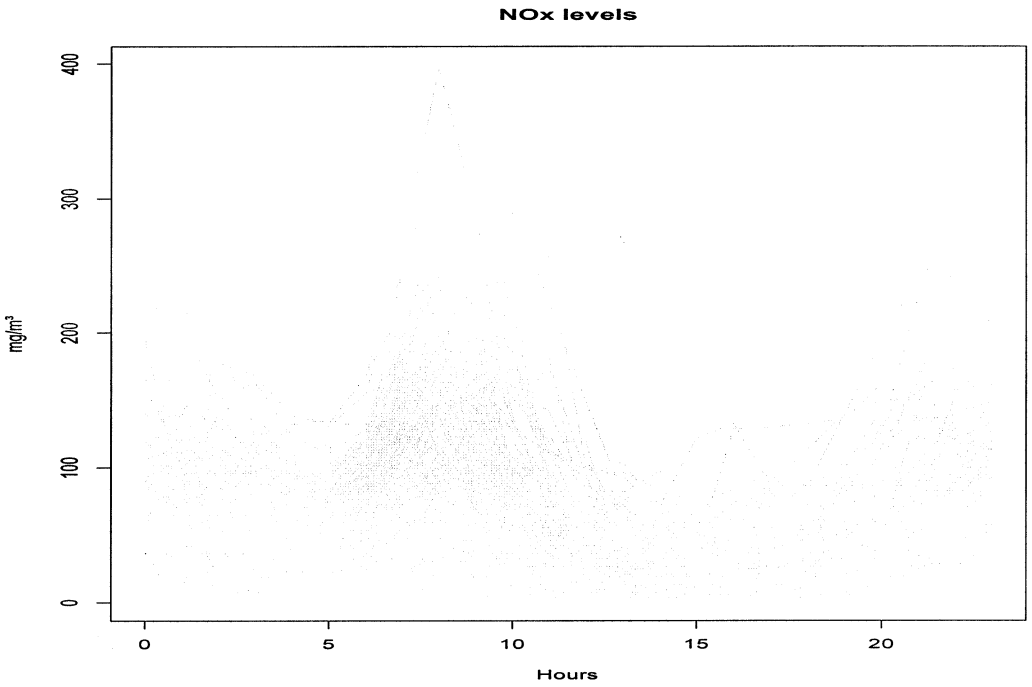


Fig. 1.4: Hourly levels of NO_x pollutants measured in Poblenou, Spain. Each curve represents one day

CHAPTER 2
TESTING FOR LACK OF DEPENDENCE IN THE FUNCTIONAL LINEAR MODEL[1]

2.1 Introduction

The last two decades have seen the emergence of new technology allowing the collection and storage of data consisting of finely sampled records over some natural repeated time or space interval. Examples include minute by minute values of a speculative asset, meteorogical and pollution monitoring data, seismic data and a plethora of examples in all fields of science and engineering. The common feature of such data is that a single observation is a curve, rather than a point or a vector. Functional Data Analysis (FDA) is a rapidly growing body of statistical tools designed to analyze such data.

One of the most popular models of the FDA is the functional linear model, see Chapters 12–17 of [4], a brief review is presented in [5]. This model is defined by the equation

$$Y_n = \Psi X_n + \varepsilon_n, \quad n = 1, 2, \ldots, N. \tag{2.1}$$

The curves Y_n and X_n, as well as the unobservable functional errors, ε_n, are assumed to lie in the Hilbert space $L^2[0,1)$. The operator $\Psi : L^2 \to L^2$ is a bounded linear operator. Detailed assumptions are stated in Section 2. Typically Ψ is assumed to be a Hilbert–Schmidt integral operator, i.e. it can be represented by a kernel

[1] Coauthored by I. Maslova, P. Kokoszka, J.J. Sojka, and L. Zhu. Reproduced by permission from Canadian Journal of Statistics, Vol. 36, No. 2, pages 207–222, 2008.

function $\psi(t,s)$ which is square integrable over $[0,1) \times [0,1)$. In that case equation (2.1) becomes

$$Y_n(t) = \int_0^1 \psi(t,s)X_n(s)ds + \varepsilon_n(t), \quad n = 1,2,\ldots,N.$$

Testing the null hypothesis of no effect, i.e testing if Ψ is zero is often a question of practical interest, which exhibits new features in the functional setting due to the fact that the data is infinitely dimensional and every dimension reduction technique restricts the domain of Ψ, and so leads to a loss of information about Ψ and complicates invertibility arguments. These issues are addressed in different contexts in [14] and [7]. The data that motivated our research requires that model (2.1) be fully functional with random explanatory variables, i.e the Y_n, X_n, ε_n are all random curves. The testing procedure we propose is similar to that developed in [7] who consider scalar responses Y_n. It turns out that the more symmetric fully functional formulation actually leads to a somewhat simpler, and more symmetric in Y_n and X_n, test statistic which does not require additional estimation of the noise variance and can be readily computed using the principal components decompositions of the the Y_n and the X_n. Our test statistic has χ^2 limiting distribution which is a good approximation for sample sizes around 50. Our asymptotic argument carefully distinguishes between population and estimated functional principal components, a point recently emphasized by [21], [22]. Other recent contributions dealing with the functional linear model are [9], [10], and [11],among others.

The research presented in this paper is to a large extent motivated by our work with magnetometer data. Figure 3.1 shows examples of magnetometer records. Technical details are explained in Section 5. Here we merely note that each panel shows one curve which we treat as a single functional observation.

The paper is organized as follows. After introducing the notation and the assumptions in Section 2, we present the test procedure and establish its asymptotic validity in Section 3. The finite sample performance is examined in Section 4, whereas the application to magnetometer data is presented in Section 5. The proofs of the asymptotic results of Section 3 are developed in Appendix A.

2.2 Notation and Assumptions

We assume that the response variables Y_n, the explanatory variables X_n and the errors ε_n are random elements of the Hilbert space $L^2[0,1)$. Recall that the expectation of a random element Z, say, of $L^2[0,1)$ is a function in $L^2[0,1)$ defined by $(\mathbf{E}Z)(t) = \mathbf{E}[Z(t)]$, $t \in [0,1)$. The inner product of $x, y \in L^2[0,1)$ is defined by $\langle x, y\rangle = \int_0^1 x(t)y(t)dt$, and the norm by $||x||^2 = \int_0^1 x^2(t)dt$. If Z is a random element of $L^2[0,1)$, then $||Z||$ is a random variable.

The theory developed below is valid under the following assumption.

ASSUMPTION 1 . The triples $(Y_n, X_n, \varepsilon_n)$ form a sequence of independent identically distributed random elements such that ε_n is independent of (Y_n, X_n) and

$$\mathbf{E}X_n = 0 \quad \text{and} \quad \mathbf{E}\varepsilon_n = 0; \tag{2.2}$$

$$\mathbf{E}||X_n||^4 < \infty \quad \text{and} \quad \mathbf{E}||\varepsilon_n||^4 < \infty. \tag{2.3}$$

Our next assumption requires that the empirical eigenelements be close to the population eigenelements of the covariance operators of the X_n and the Y_n. This point is often overlooked in empirical work, but assumptions of this type are needed

to develop a rigorous asymptotic theory, see Chapter 4 of [23] and [21].

Introduce the operators:

$$\Gamma x = \mathbf{E}[\langle X_1, x\rangle X_1], \quad \Lambda x = \mathbf{E}[\langle Y_1, x\rangle Y_1], \quad \Delta x = \mathbf{E}[\langle X_1, x\rangle Y_1].$$

Denote their empirical counterparts by $\Gamma_N, \Lambda_N, \Delta_N$, e.g.

$$\Gamma_N x = \frac{1}{N}\sum_{n=1}^{N} \langle X_n, x\rangle X_n.$$

Define the eigenelements by

$$\Gamma v_k = \gamma_k v_k, \quad \Lambda u_j = \lambda_j u_j.$$

Empirical eigenelements are defined correspondingly and denoted by $(\hat{\gamma}_k, \hat{v}_k), (\hat{\lambda}_j, \hat{u}_j)$.

ASSUMPTION 2 . The eigenvalues of the operators Γ and Λ satisfy, for some $p > 0$ and $q > 0$,

$$\gamma_1 > \gamma_2 > \ldots \gamma_p > \gamma_{p+1}, \quad \lambda_1 > \lambda_2 > \ldots \lambda_q > \lambda_{q+1}. \tag{2.4}$$

Assumption 2 implies that the eigenspaces corresponding to the first largest p (respectively q) eigenvalues are one dimensional. Therefore, the corresponding normalized principal components are well–defined (up to the sign) and orthogonal. No formal test is currently available to verify Assumption 2, but it is very natural as in applications the estimated eigenvalues are always positive and distinct.

In the proofs, we will often use the relations

$$\limsup_{N\to\infty} N\mathbf{E}||v_k - \hat{v}_k||^2 < \infty, \quad \limsup_{N\to\infty} N\mathbf{E}||u_j - \hat{u}_j||^2 < \infty; \tag{2.5}$$

$$\limsup_{N\to\infty} N\mathbf{E}\left[|\gamma_k - \hat{\gamma}_k|^2\right] < \infty, \quad \limsup_{N\to\infty} N\mathbf{E}\left[|\lambda_j - \hat{\lambda}_j|^2\right] < \infty, \tag{2.6}$$

which hold for each $k \leq p$ and $j \leq q$ under Assumptions 1 and 2, see Chapter 4 of [23].

2.3 Test Procedure and Asymptotic Results

Assuming model (2.1), we wish to test

$$H_0 : \Psi = 0 \quad \text{versus} \quad H_A : \Psi \neq 0.$$

We thus test the null hypothesis that the curves X_n have no effect on the curves Y_n. This is analogous to testing in the scalar linear model, $y_i = \beta_0 + \beta_1 x_{i,1} + \ldots \beta_{p-1} x_{i,p-1} + \varepsilon_i$, whether the regression on the regressors $X_1, \ldots X_{p-1}$ is significant. In this standard setting, the F–test is used. In the particular case of straight line regression, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the F–test is equivalent to the usual t–test for non–zero slope, see e.g. Chapter 4 of [24]. In our functional setting the slope corresponds to a linear operator which transforms functions into functions. Just as in the case of straight line regression, the nullity of Ψ does not mean that there is no dependence between the curves X_n and Y_n, but that if there is a dependence, it cannot be described by a functional linear model.

The testing procedure involves restrictions of the operators defined in Section 2 to certain finite dimensional subspaces. This is a dimension reduction procedure which necessarily involves some loss of information about the action of Ψ. The subspace

$\mathcal{V}_p = \text{sp}\{v_1, \ldots, v_p\}$, which is isomorphic to R^p, contains the best approximations to the X_n which are linear combinations of the first p principal components, see Section 8.2.3 of [4]. Similarly, $\mathcal{U}_q = \text{sp}\{u_1, \ldots, u_q\}$ is a good approximation to $\text{sp}\{Y_1, \ldots, Y_n\}$.

Since, by (2.1), $\Delta = \Psi\Gamma$, we have, for $k \leq p$,

$$\Psi v_k = \gamma_k^{-1} \Delta v_k. \tag{2.7}$$

Thus, by Assumption 2, Ψ vanishes on $\text{sp}\{v_1, \ldots, v_p\}$ if and only if $\Delta v_k = 0$ for each $k = 1, \ldots, p$. Observe that

$$\Delta v_k \approx \Delta_N v_k = \frac{1}{N} \sum_{n=1}^{N} \langle X_n, v_k \rangle Y_n.$$

Since $\text{sp}\{Y_1, \ldots, Y_N\}$ is well approximated by $\mathcal{U}_q$, a test can be developed by checking if

$$\langle \Delta_N v_k, u_j \rangle = 0, \quad k = 1, \ldots, p, \ j = 1, \ldots, q. \tag{2.8}$$

If such a test accepts H_0, it means that for every $x \in \mathcal{V}_p$, Ψx is not in $\mathcal{U}_q$. Intuitively, it means that up to a small error arising from the approximations by the principal components and a random error, no function Y_n, $n = 1, 2, \ldots, N$, can be expressed as a linear combination of functions X_n, $n = 1, 2, \ldots, N$.

Theorem 1 shows that test statistic

$$\hat{T}_N(p, q) = N \sum_{k=1}^{p} \sum_{j=1}^{q} \hat{\gamma}_k^{-1} \hat{\lambda}_j^{-1} \langle \Delta_N \hat{v}_k, \hat{u}_j \rangle^2 \tag{2.9}$$

has a parameter–free asymptotic distribution.

THEOREM 1. *Under H_0 and the assumptions of Section 2,*

$$\hat{T}_N(p,q) \xrightarrow{d} \chi^2_{pq}.$$

If H_0 fails, then $\Psi v_k \neq 0$ for some $k \geq 1$. If we impose conditions only on the first p largest eigenvalues, the test will be consistent only if Ψ does not vanish on one of the v_k, $k = 1, 2, \ldots, p$. The test has no power if Ψ does not vanish on the orthogonal complement of $\mathrm{sp}\{v_1, \ldots, v_p\}$. Further, to ensure consistency, one of the v_k, $k = 1, 2, \ldots, p$ must be mapped into $\mathrm{sp}\{u_1, \ldots, u_q\}$. These restrictions are intuitively appealing because we want to test if the main sources of the variability of the responses Y can be explained by the main sources of the variability of the explanatory variables X.

The following theorem formalizes these ideas and establishes the consistency of the test.

THEOREM 2. *If the assumptions of Section 2 hold, and $\langle \Psi v_k, u_j \rangle \neq 0$ for some $k \leq p$ and $j \leq q$, then $\hat{T}_N(p,q) \xrightarrow{P} \infty$, as $N \to \infty$.*

In linear regression setting, it is often of interest to test if specific covariates have no effect on the responses. In our setting, we could ask if specific principal components v_k have no effect. It is easy to see from the proof of Theorem 1, see Lemma 1 in particular, that if we want to test if principal components $v_{i(1)}, \ldots, v_{i(p')}$ have no effect, we must modify the statistic (2.9) by including only these components. The limit χ^2 distribution will then have $p'q$ degrees of freedom. A further obvious modification can be made if we want to check if there is an effect in the subspace spanned by some principal components of the responses Y_k. Modifications of this

type are useful if some principal components have obvious physical interpretations. This is sometimes the case in space physics applications, see [18], but in the case of when the X_n are high-latitude records, see Section 5, the v_k cannot, at this point, be readily interpreted.

Summary of the testing procedure.

1. Check the linearity assumption using FPC score predictor-response plots, see Section 5.

2. Select the number of important PC's, p and q using both the scree test and CPV, see Section 5.

3. Compute the test statistics $\hat{T}_N(p,q)$ (2.9). Note that

$$\langle \triangle_N \hat{v}_k, \hat{u}_j \rangle = \langle \frac{1}{N} \sum_{n=1}^{N} \langle X_n, \hat{v}_k \rangle Y_n, \hat{u}_j \rangle = \frac{1}{N} \sum_{n=1}^{N} \langle X_n, \hat{v}_k \rangle \langle Y_n, \hat{u}_j \rangle,$$

where $\langle X_n, \hat{v}_k \rangle$ is the kth score of the X_n, and $\langle Y_n, \hat{u}_j \rangle$ is jth score of the Y_n. These scores and the eigenvalues $\hat{\gamma}_k$ and $\hat{\lambda}_j$ are output of functions available in the `R` package `fda`.

4. If $\hat{T}_N(p,q) > \chi^2_{pq}(\alpha)$, reject the null hypothesis of no linear effect. The critical value $\chi^2_{pq}(\alpha)$ is the $(1-\alpha)$th quantile of the chi-squared distribution with pq degrees of freedom.

2.4 A Small Simulation Study

In this section, we present the results of a small simulation study intended to evaluate the empirical size and power of the test in standard Gaussian settings.

We used $R = 1000$ replications of samples of processes ε_n, X_n and Y_n, $n = 1, 2, \ldots, N$. In order to evaluate the empirical size, we generated samples of pairs (ε_n, Y_n) with independent components. To find the empirical power, we generated

samples of pairs (ε_n, X_n) with independent components, and calculated Y_n according to (2.1). As ε_n, X_n and Y_n, we used Brownian bridge and motion processes in various combinations. The computations were performed using the `R` package `fda`. We used both Fourier and splines bases.

Since the Brownian bridge and motion have very regular Karhunene-Loeve decompositions, see e.g. [23], p. 26, it is not surprising that the size and power of the test do not depend appreciably on p and q. Figures 2.2 and 2.3 illustrate this point. The horizontal axes represent various combinations of p and q; 1 stands for $p = 1$ and $q = 1$; 2 for $p = 1$, $q = 2$; 3 for $p = 1$, $q = 3$, etc. All combinations of $p \leq 4$, $q \leq 4$ were considered in the size study and $p \leq 6, q \leq 6$ in the power study. The results for Brownian bridges and motions and Fourier and spline bases are practically the same. For this reason, we present the results only in cases when all processes are Brownian bridges, and the analysis was performed with the Fourier basis.

Naturally, the bigger the sample size the closer the empirical size of the test is to the nominal size. Nevertheless, there is little or no improvement in the size of the test starting from $N = 40$ — 80; these values can therefore be considered sufficient to obtain reasonable size; with $N = 40$ the test being slightly conservative.

To evaluate the empirical power, we used the Gaussian kernel

$$\psi(s,t) = C \exp(t^2 + s^2)/2, \qquad t \in [0,1], s \in [0,1] \tag{2.10}$$

with constants C such that $||\Psi|| < 1$, i.e. $|C| < 1$. Panels (a) and (b) of Figure 2.3 present power when the dependence between X_n and Y_n is quite strong, $||\Psi|| = 0.75$. For $N = 80$, the power is practically 100% if $||\Psi|| = 0.75$. The right column of Figure 2.3 shows the power of the test when $||\Psi|| = 0.5$. In this case power increases slower with N.

Even though this paper is concerned with testing for no effect in the *linear* fully functional model, it might be interesting to see what happens if the responses depend or the regressors in a nonlinear manner. Let X_n be independent Brownian motions, and ε_n independent Brownian bridges (independent of the X_n). We computed the empirical power of the test for the following models:

$$Y_n(t) = H_2(X_n(t)) + \varepsilon_n(t), \tag{2.11}$$

where $H_2(x) = x^2 - 1$ and

$$Y_n(t) = X_n(t)\varepsilon_n(t). \tag{2.12}$$

The function H_2 in (2.11) is the Hermite polynomial of rank 2; in model (2.12) the errors are multiplicative.

For $N = 40$, in case of model (2.11), the empirical power for various principal component combinations is around 53% for the significance level $\alpha = 10\%$, 30% for $\alpha = 5\%$, and 9% for $\alpha = 1\%$. For the multiplicative model (2.12) the power is about 38% for $\alpha = 10\%$, 24% for $\alpha = 5\%$, and 6% for $\alpha = 1\%$. Just as in the case of usual linear models, the test can detect some nonlinear dependence, but not reliably.

2.5 Application to Magnetometer Data

About a hundred terrestrial geomagnetic observatories form a network, INTERMAGNET, designed to monitor and understand the behavior of electrical currents flowing in the magnetosphere and ionosphere (M-I). Interestingly, C. F. Gauss was one of the leaders of the early nineteenth century effort to establish such a network, and pioneered the statistical analysis of the resulting measurements; see Chapter 1 of [2]. Modern digital magnetometers record three components of the magnetic

field in five second resolution, but the data made available by INTERMAGNET (http://www.intermagnet.org) consist of one minute averages (1440 data points per day per component per observatory). Figure 3.1 shows examples of magnetometer records. We work with the Horizontal (H) component of the magnetic field. This is the component lying in the Earth's tangent plane and pointing toward the magnetic North. It most directly reflects the variation of the M-I currents we wish to study. The M-I currents form a complex interactive system which at present is only partially understood, see [25] and [26]. The magnetometer records contain intertwined signatures of many currents, and an effort has been under way to deconvolute the signatures of various currents. So far this has been done by preprocessing records from every individual station, and then combining the filtered signals from stations at the same magnetic latitude (e.g. equatorial stations, or auroral stations), see [19] for a recent example of such an approach. It is however believed, see e.g. [17], that the auroral currents may have, a perhaps indirect, impact on the equatorial and mid-latitude currents. The present paper has been motivated by this problem and shows that this is indeed the case using the proposed test of significance. Our goal in this section is to illustrate the methodology using an interesting data set rather than to present a comprehensive case study. A detailed analysis with a deeper discussion of physical insights is presented in [27].

The data consist of minute–by–minute records of the horizontal intensity of the magnetic field measured in 2001 at observatories listed in Table 5.1. The observatories in each of the four groups are roughly aligned along the same magnetic longitude. The functional observations consist of daily (in UT) curves (1440 records per curve). Examples of such curves are shown in Figure 3.1.

The question of interest is whether the auroral geomagnetic activity reflected in the high–latitude curves has an effect on the processes in the equatorial belt reflected

by the mid– and high– latitude curves. This question is of particular interest for days during which a high–latitude activity known as a substorm occurs. Its most spectacular manifestation are the Norther Lights caused by high–energy electrojets flowing for a few hours in the auroral belt. The top left panel of Figure 3.1 shows a signature of a substorm. It is believed that there is energy transfer between the auroral electrojets and lower latitude currents, but the direct physical mechanisms which might be responsible for this interaction are a matter of debate.

The question can be cast into the setting of the functional linear model (2.1) in which the X_n are centered high–latitude records and Y_n are centered mid– or low–latitude records. This postulates an approximate statistical model for the data and allows us to the the null hypothesis $\Psi = 0$. If the null is true, we conclude that the high–latitude curves X_n have no linear effect on the lower latitude curves. If the null is rejected, this indicates the existence of an effect, which can be approximately linear (in the functional sense). Other modeling settings are conceivable, an adaptation of a nonparametric approach advocated by [28] might be appealing and could provide additional insights.

In the analysis below, we use $N = 41$ days in January – August 2001 which contained a medium strength substorm. Thus X_n is the curve on the nth day with a medium strength substorm and Y_n is the curve on the same (UT) day measured at mid– or low–latitude station. In addition, we consider mid– and low–latitude curves 1, 2 and 3 days after the day with a substorm. This is intended to check how long the effects of a substrom persist. The independence of the cases (X_n, Y_n) can be assumed to hold approximately because the substorm days are typically separated by quiet days during which the M–I system resets itself. The independence of the X_n is also confirmed by the application of the test developed by [29]. The same holds true for the Y_n.

As mentioned in Section 4, to ensure that the test gives reliable results, linearity assumption must be checked. For this purpose, visual techniques introduced by [12] can be used. Functional principal component (FPC) scores are used to check the linearity assumption. In case of linear dependence, the FPC score plots are roughly football–shaped. When the dependence is not linear, these plots exhibit different patterns. For example, for model (2.11) introduced in Section 4, the scatterplot of the first FPC clearly shows a quadratic trend, see Figure 3.3.

Figure 3.4 is an example of the relationship between the response and the predictor FPC scores for magnetometer data. We used CMO records as X, and THY with no lag – as Y. These scatterplots indicate linear relationship with some outliers. Since we do not require Gaussianity, only finite fourth moment, these outliers need not invalidate our conclusions. In case of other pairs of functional data, the FPC score plots look similar. We conclude that a linear model is approximately appropriate for our application.

To apply the test, we need to decide which values of p and q should be used. We propose to use all values up to some meaningful upper bounds, and look at the pattern of rejections and acceptances as a function of p and q. One of the ways to pick the most important principal components is to use the scree test, which is a graphical method first proposed by [30]. To apply the scree method one plots the successive eigenvalues against the corresponding principal components (see Figure 3.5). The method suggests to find the place where the smooth decrease of eigenvalues appears to level off. To the right of this point one finds only "factorial scree" ("scree" is a geological term referring to the debris which collects on the lower part of a rocky slope).

Another way to pick the unknown number of principal components from the data is to compute the cumulative percentage of total variance (CPV), as in multi-

variate principal component analysis. So, the CPV explained by the first p functional principal components is

$$CPV(p) = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{\infty} \lambda_k}.$$

Table 2.2 gives the upper limits on p and q together with the CPV explained by these components. Visual examination of the principal components beyond the upper bound confirms that they resemble a random noise.

In most cases, there is a clear rejection or acceptance for almost all combinations of p and q, at for all small values which correspond to the most important principal components. For such cases, we can with reasonable confidence reject ("1") or fail to reject ("0"). However, there are some cases where it is not clear what conclusion to draw. We denote them by "1?" – inclined toward rejecting H_0, "0?"– inclined toward accepting H_0, "1?0?"– inconclusive. Figure 3.7 gives examples of such cases.

Table 2.3 presents the results of our analysis. It shows that the effect of a substorm persist for about one day. Beyond that time, the magnetometer data at mid– and low latitudes are not linearly (in the functional sense) dependent on the high latitude records. We note however that a slightly more complex picture emerges for different seasons in 2001 and for special subcategories of substorms. These issues are discussed in [27].

2.6 Proofs of Theorem 1 and 2

Proof of Theorem 1. Theorem 1 follows from Corollary 1, which is arrived at through a series of lemmas. Lemma 1 shows that the χ^2 limit holds for the population eigenelements. The remaining lemmas show that the differences between the empirical and population eigenelements have asymptotically negligible effect.

LEMMA 1. *Under H_0 and the assumptions of Section 2, for each $j \leq q, k \leq p$,*

$$\sqrt{N}\,\langle \Delta_N v_k, u_j \rangle \xrightarrow{d} \eta_{kj}\sqrt{\gamma_k \lambda_j}, \tag{2.13}$$

with $\eta_{kj} \sim N(0,1)$. Moreover, η_{kj} and $\eta_{k'j'}$ are independent if $(k,j) \neq (k',j')$.

Proof of Lemma 1. Under H_0,

$$\sqrt{N}\,\langle \Delta_N v_k, u_j \rangle = N^{-1/2} \sum_{n=1}^{N} \langle X_n, v_k \rangle \langle \varepsilon_n, u_j \rangle .$$

The summands have mean zero and variance $\gamma_k \lambda_j$, so (2.13) follows.

To verify that η_{kj} and $\eta_{k'j'}$ are independent if $(k,j) \neq (k',j')$, it suffices to show that $\sqrt{N}\,\langle \Delta_N v_k, u_j \rangle$ and $\sqrt{N}\,\langle \Delta_N v_{k'}, u_{j'} \rangle$ are uncorrelated. Observe that

$$\mathbf{E}\left[\sqrt{N}\,\langle \Delta_N v_k, u_j \rangle, \sqrt{N}\,\langle \Delta_N v_{k'}, u_{j'} \rangle\right]$$

$$= \frac{1}{N}\mathbf{E}\left[\sum_{n=1}^{N} \langle X_n, v_k \rangle \langle \varepsilon_n, u_j \rangle \sum_{n'=1}^{N} \langle X_{n'}, v_{k'} \rangle \langle \varepsilon_{n'}, u_{j'} \rangle\right]$$

$$= \frac{1}{N}\sum_{n,n'=1}^{N} \mathbf{E}\left[\langle X_n, v_k \rangle \langle X_{n'}, v_{k'} \rangle\right] \mathbf{E}\left[\langle \varepsilon_n, u_j \rangle \langle \varepsilon_{n'}, u_{j'} \rangle\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N} \mathbf{E}\left[\langle X_n, v_k \rangle \langle X_n, v_{k'} \rangle\right] \mathbf{E}\left[\langle \varepsilon_n, u_j \rangle \langle \varepsilon_n, u_{j'} \rangle\right]$$

$$= \langle \Gamma v_k, v_{k'} \rangle \langle \Lambda u_j, u_{j'} \rangle = \lambda_k \delta_{kk'} \lambda_j \delta_{jj'}.$$

Recall that the Hilbert–Schmidt norm of a Hilbert–Schmidt operator S is defined by

$$||S||_S^2 = \sum_{j=1}^{\infty} ||Se_j||^2,$$

where $\{e_1, e_2, \ldots\}$ is any orthonormal basis. Recall also that the Hilbert–Schmidt norm dominates the operator norm: $||S|| \leq ||S||_{\mathcal{S}}$.

LEMMA 2. *Under H_0 and the assumptions of Section 2,*

$$\boldsymbol{E}||\Delta_N||_{\mathcal{S}}^2 = N^{-1}\boldsymbol{E}||X_1||^2\boldsymbol{E}||\varepsilon_1||^2.$$

Proof of Lemma 2. Observe that

$$||\Delta_N e_j||^2 = N^{-2}\sum_{n,n'=1}^{N}\langle X_n, e_j\rangle\langle X_{n'}, e_j\rangle\langle Y_n, Y_{n'}\rangle.$$

Therefore, under H_0,

$$\mathbf{E}||\Delta_N||_{\mathcal{S}}^2 = N^{-2}\sum_{j=1}^{\infty}\sum_{n,n'=1}^{N}\mathbf{E}\left[\langle X_n, e_j\rangle\langle X_{n'}, e_j\rangle\langle \varepsilon_n, \varepsilon_{n'}\rangle\right]$$

$$= N^{-2}\sum_{j=1}^{\infty}\sum_{n=1}^{N}\mathbf{E}\langle X_n, e_j\rangle^2\mathbf{E}||\varepsilon_n||^2 = N^{-1}\mathbf{E}||\varepsilon_1||^2\sum_{j=1}^{\infty}\langle X_1, e_j\rangle^2 = N^{-1}\mathbf{E}||\varepsilon_1||^2\mathbf{E}||X_1||^2.$$

The following elementary lemma is stated for ease of reference.

LEMMA 3. *Suppose $\{U_N\}$ and $\{V_N\}$ are random sequences in a Hilbert space such that $||U_N|| \xrightarrow{P} 0$ and $||V_N|| = O_P(1)$ i.e. $\lim_{C\to\infty}\limsup_{N\to\infty}P(||V_N|| > C) = 0$. Then*

$$\langle U_N, V_N\rangle \xrightarrow{P} 0.$$

Proof of Lemma 3. The Lemma follows from the corresponding property of real random sequences and the inequality $|\langle U_N, V_N\rangle| \leq ||U_N||||V_N||$.

LEMMA 4. *Under H_0 and the assumptions of Section 2, for each $j \leq q, k \leq p$,*

$$\sqrt{N}\langle \Delta_N \hat{v}_k, \hat{u}_j\rangle \xrightarrow{d} \eta_{kj}\sqrt{\gamma_k \lambda_j}, \tag{2.14}$$

with η_{kj} equal to those in Lemma 1 .

Proof of Lemma 4. It suffices to verify that

$$\sqrt{N}\langle \Delta_N \hat{v}_k, \hat{u}_j\rangle - \sqrt{N}\langle \Delta_N v_k, u_j\rangle \xrightarrow{P} 0. \tag{2.15}$$

Relation (2.15), will follow from

$$\sqrt{N}\langle \Delta_N v_k, \hat{u}_j - u_j\rangle \xrightarrow{P} 0 \tag{2.16}$$

and

$$\sqrt{N}\langle \Delta_N (\hat{v}_k - v_k), \hat{u}_j\rangle \xrightarrow{P} 0. \tag{2.17}$$

To verify (2.16), note that by (2.5), $\sqrt{N}(\hat{u}_j - u_j) = O_P(1)$, and by Lemma 2, $\mathbf{E}||\Delta_N v_k|| \leq \mathbf{E}||\Delta_N||_S = O(N^{-1/2})$. Thus (2.16) follows from Lemma 3.

To use the same argument for (2.17) (with (2.5)), we note that

$$\sqrt{N}\langle \Delta_N (\hat{v}_k - v_k), \hat{u}_j\rangle = \sqrt{N}\left\langle \hat{v}_k - v_k, \tilde{\Delta}_N \hat{u}_j\right\rangle,$$

where $\tilde{\Delta}_N x = N^{-1} \sum_{n=1}^{N} \langle Y_n, x \rangle X_n$. Lemma 2 shows that under H_0, $\mathbf{E}||\tilde{\Delta}_N||_S = \mathbf{E}||\Delta_N||_S$.

By (2.6), $\hat{\gamma}_k \xrightarrow{P} \gamma_k$ and $\hat{\lambda}_j \xrightarrow{P} \lambda_j$, so we obtain

COROLLARY 1. *Under H_0 and the assumptions of Section 2, for each $j \leq q, k \leq p$,*

$$\sqrt{N}\hat{\gamma}_k^{-1/2}\hat{\lambda}_j^{-1/2} \langle \Delta_N \hat{v}_k, \hat{u}_j \rangle \xrightarrow{d} \eta_{kj}, \tag{2.18}$$

with η_{kj} equal to those in Lemma 1.

Proof of Theorem 2. Denote

$$\hat{S}_N(p,q) = \sum_{k=1}^{p} \sum_{j=1}^{q} \hat{\gamma}_k^{-1} \hat{\lambda}_j^{-1} \langle \Delta_N \hat{v}_k, \hat{u}_j \rangle^2 .$$

By Lemma 7 and (2.6), $\hat{S}_N(p,q) \xrightarrow{P} S(p,q) > 0$. Hence $\hat{T}_N(p,q) = N\hat{S}_N(p,q) \xrightarrow{P} \infty$.

To establish Lemma 7, it is convenient to split the argument into two simple lemmas: Lemma 5 and Lemma 6.

LEMMA 5. *If Y_n, $n \geq 1$, are identically distributed, then $\boldsymbol{E}||\Delta_N|| \leq \boldsymbol{E}||Y_1||^2$.*

Proof of Lemma 5. For arbitrary $u \in L^2$ with $||u|| \leq 1$,

$$||\Delta_N u|| \leq N^{-1} \sum_{n=1}^{N} |\langle Y_n, u \rangle| ||Y_n|| \leq N^{-1} \sum_{n=1}^{N} ||Y_n||^2.$$

Since the Y_n are identically distributed, the claim follows.

LEMMA 6. *Under the assumptions of Section 2, for any functions* $v, u \in L^2$,

$$\langle \Delta_N v, u \rangle \xrightarrow{P} \langle \Delta v, u \rangle .$$

Proof of Lemma 6. The result follows from the Law of Large Numbers after noting that

$$\langle \Delta_N v, u \rangle = \frac{1}{N} \sum_{n=1}^{N} \langle X_n, v \rangle \langle Y_n, u \rangle$$

and

$$\mathbf{E}\left[\langle X_n, v \rangle \langle Y_n, u \rangle\right] = \mathbf{E}\left[\langle \langle X_n, v \rangle Y_n, u \rangle\right] = \langle \Delta v, u \rangle .$$

LEMMA 7. *Under the assumptions of Section 2,* $\langle \Delta_N \hat{v}_k, \hat{u}_j \rangle \xrightarrow{P} \langle \Delta v_k, u_j \rangle$, $j \leq q, k \leq p$.

Proof of Lemma 7. By Lemma 6, it suffices to show

$$\langle \Delta_N v_k,\ \hat{u}_j - u_j \rangle \xrightarrow{P} 0; \tag{2.19}$$

$$\langle \Delta_N \hat{v}_k - \Delta_N v_k,\ \hat{u}_j \rangle \xrightarrow{P} 0. \tag{2.20}$$

These relations follow from Lemma 3, relations (2.6) and Lemma 5.

Table 2.1: Geomagnetic observatories used in this study.

Latitude	I	II	III	IV
High	College (CMO)	–	–	–
Mid	Boulder (BOU)	Fredericksburg (FRD)	Tihany (THY)	Memambetsu (MMB)
Low	Honolulu (HON)	San Juan (SJG)	Hermanus (HER)	Kakioka (KAK)

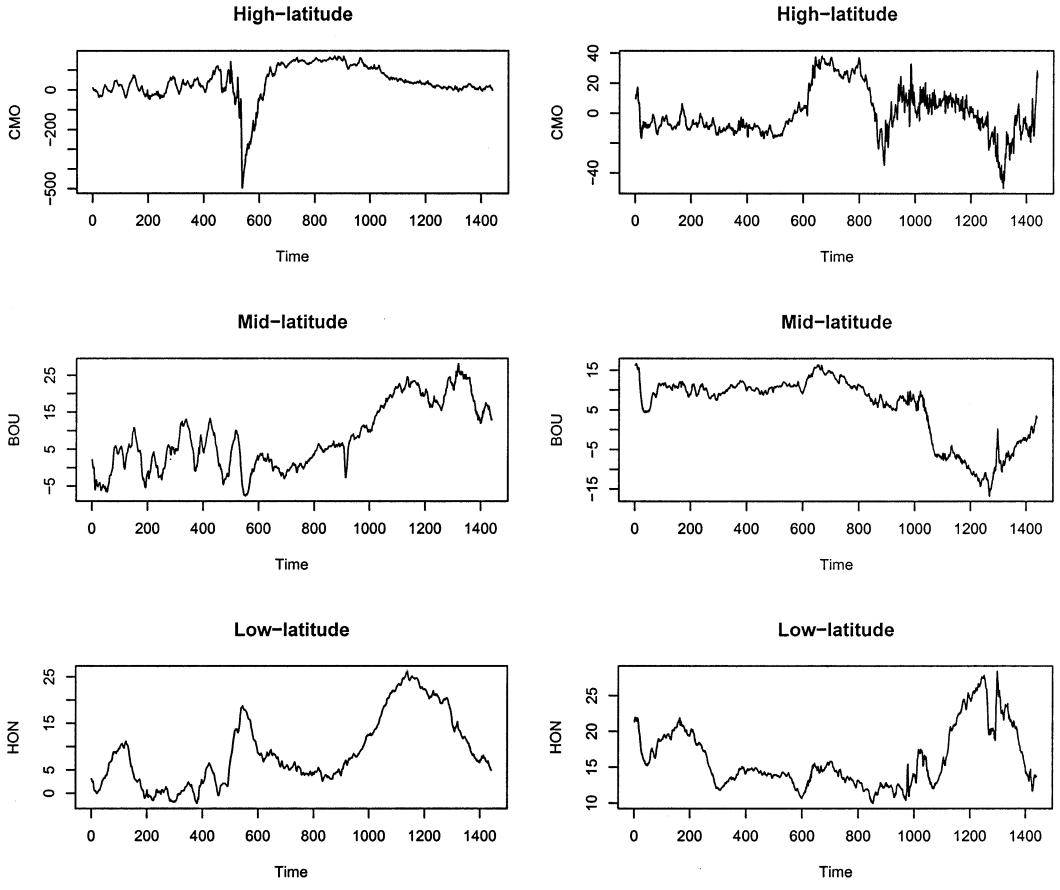


Fig. 2.1: Horizontal intensities of the magnetic field measured at a high-, mid- and low-latitude stations during a sub-storm (left column) and a quiet day (right column). Note the different vertical scales for high-latitude records.

Table 2.2: Number of principal components retained by the scree test, and percentage of total variability explained, during medium strength sub-storm days that occurred from January until August, 2001.

Stations	PC	%
College (CMO)	10	81.97
Boulder (BOU)	3	86.99
Boulder (BOU) one-day lag	4	81.68
Boulder (BOU) two-day lag	2	91.18
Boulder (BOU) three-day lag	3	95.15
Honolulu (HON)	2	93.85
Honolulu (HON) one-day lag	4	93.89
Honolulu (HON) two-day lag	3	98.16
Honolulu (HON) three-day lag	2	98.80
Fredericksburg (FRD)	4	92.83
Fredericksburg (FRD) one-day lag	4	89.52
Fredericksburg (FRD) two-day lag	3	94.35
Fredericksburg (FRD) three-day lag	4	96.77
San Juan (SJG)	2	90.86
San Juan (SJG) one-day lag	3	86.40
San Juan (SJG) two-day lag	2	94.32
San Juan (SJG) three-day lag	3	96.63
Tihany (THY)	3	89.57
Tihany (THY) one-day lag	4	83.75
Tihany (THY) two-day lag	2	89.64
Tihany (THY) three-day lag	3	94.33
Hermanus (HER)	2	90.91
Hermanus (HER) one-day lag	3	89.64
Hermanus (HER) two-day lag	2	93.84
Hermanus (HER) three-day lag	3	96.53
Memambetsu (MMB)	2	89.99
Memambetsu (MMB) one-day lag	3	85.36
Memambetsu (MMB) two-day lag	3	95.64
Memambetsu (MMB) three-day lag	3	97.04
Kakioka (KAK)	2	92.89
Kakioka (KAK) one-day lag	2	92.89
Kakioka (KAK) two-day lag	3	97.08
Kakioka (KAK) three-day lag	3	98.04

Table 2.3: Results of the test for medium strength sub-storm days that occurred from January to August, 2001.

CMO							
BOU0	BOU1	BOU2	BOU3	HON0	HON1	HON2	HON3
1?	0	0	0?	1?	1?0?	0	0?
FRD0	FRD1	FRD2	FRD3	SJG0	SJG1	SJG2	SJG3
1?	0	0	0?	1?	0	0	0?
THY0	THY1	THY2	THY3	HER0	HER1	HER2	HER3
1?	0?	0	0?	0?	0?	0	0?
MMB0	MMB1	MMB2	MMB3	KAK0	KAK1	KAK2	KAK3
0?	0	0	0?	1?	1?	0	0?

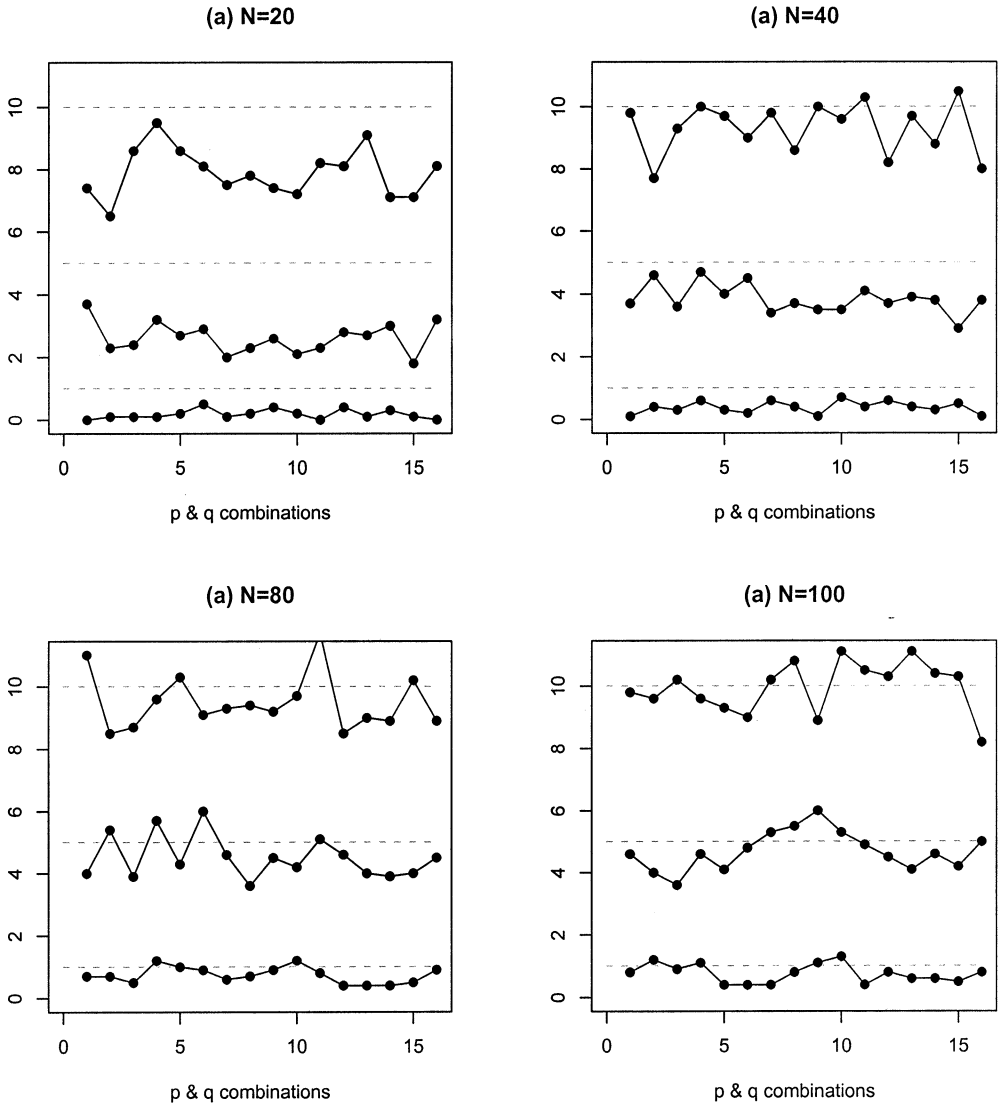


Fig. 2.2: Empirical size of the test for $\alpha = 1\%, 5\%, 10\%$ (indicated by dotted lines) for different combinations of p and q. Here ε_n and Y_n, $n = 1, 2, \ldots, N$ are two independent Brownian Bridges.

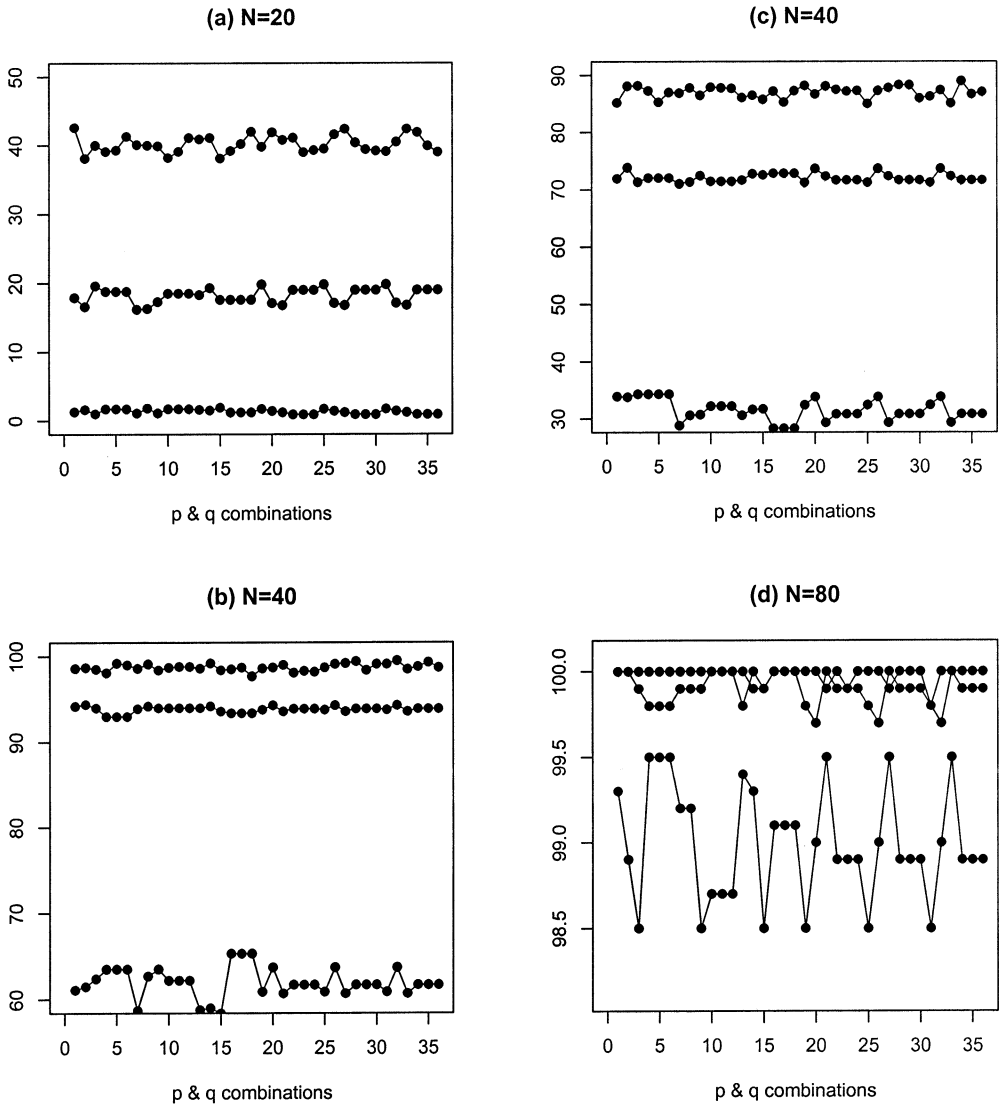


Fig. 2.3: Empirical power of the test for different combinations of principal components and different sample sizes N. Here X_n and ε_n are Brownian Bridges. In panels (a), (b) $||\Psi|| = 0.75$; in panels (c), (d) $||\Psi|| = 0.5$.

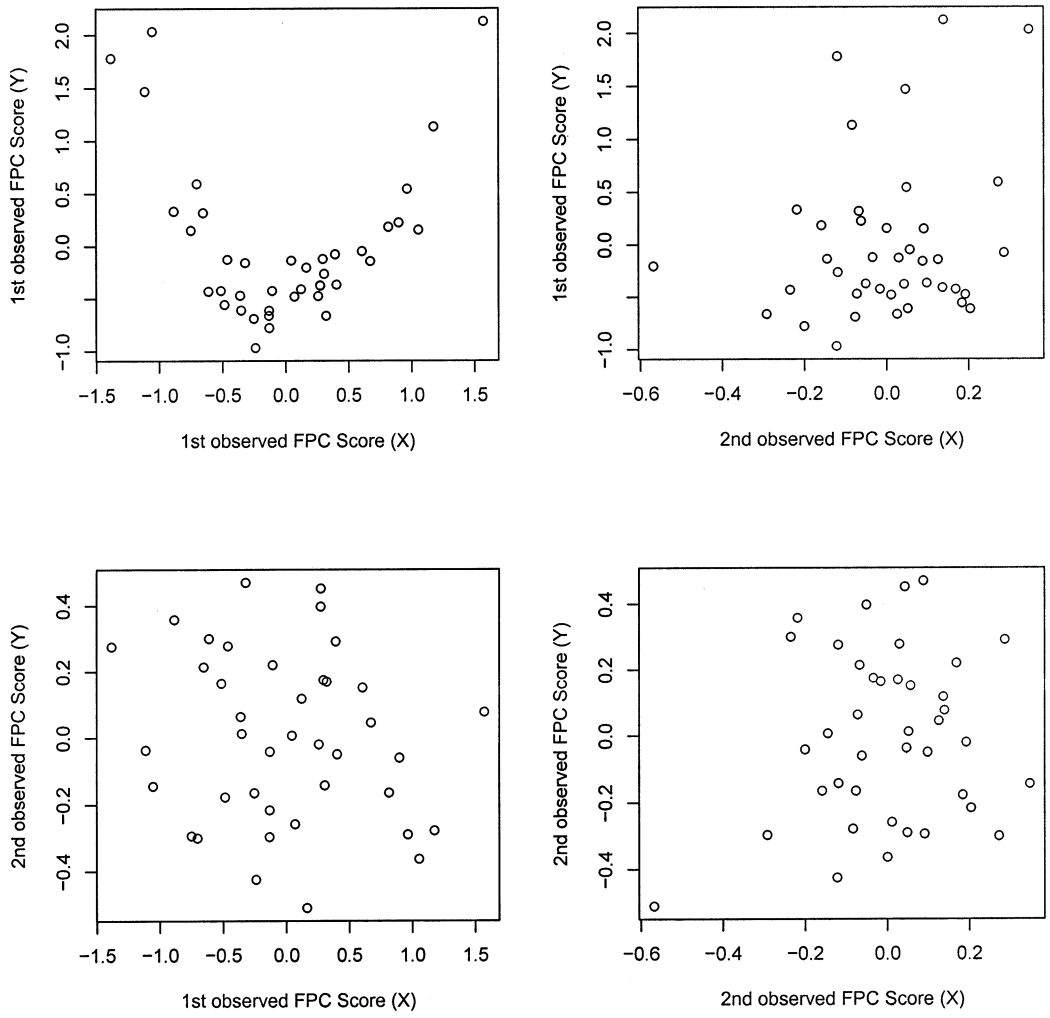


Fig. 2.4: Functional predictor-response plots of functional principal component scores of response functions versus functional principal component scores of predictor functions for $Y_n(t) = H_2(X_n(t)) + \varepsilon_n(t)$, where $H_2(x) = x^2 - 1$, $n = 1, \ldots, 40$.

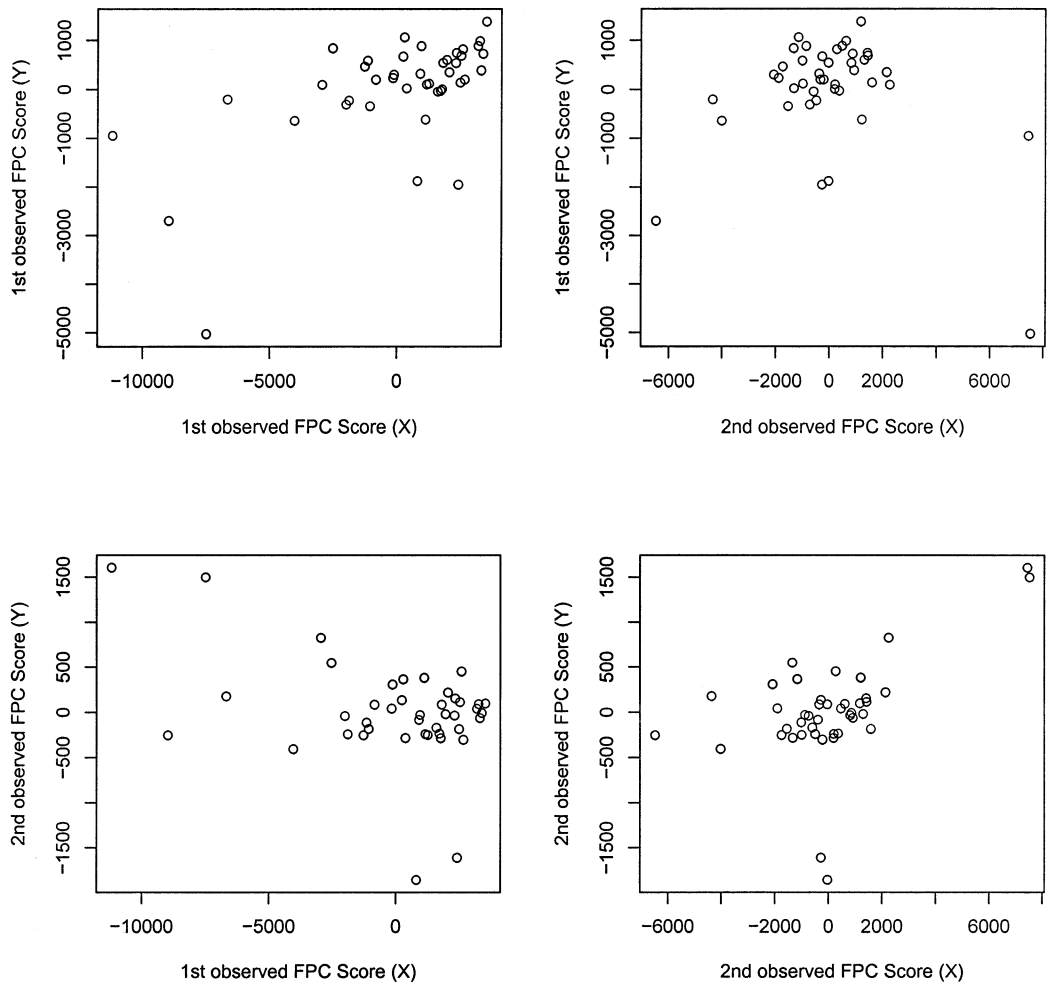


Fig. 2.5: Functional predictor-response plots of functional principal component scores of response functions versus functional principal component scores of predictor functions for magnetometer data (CMO vs THY0)

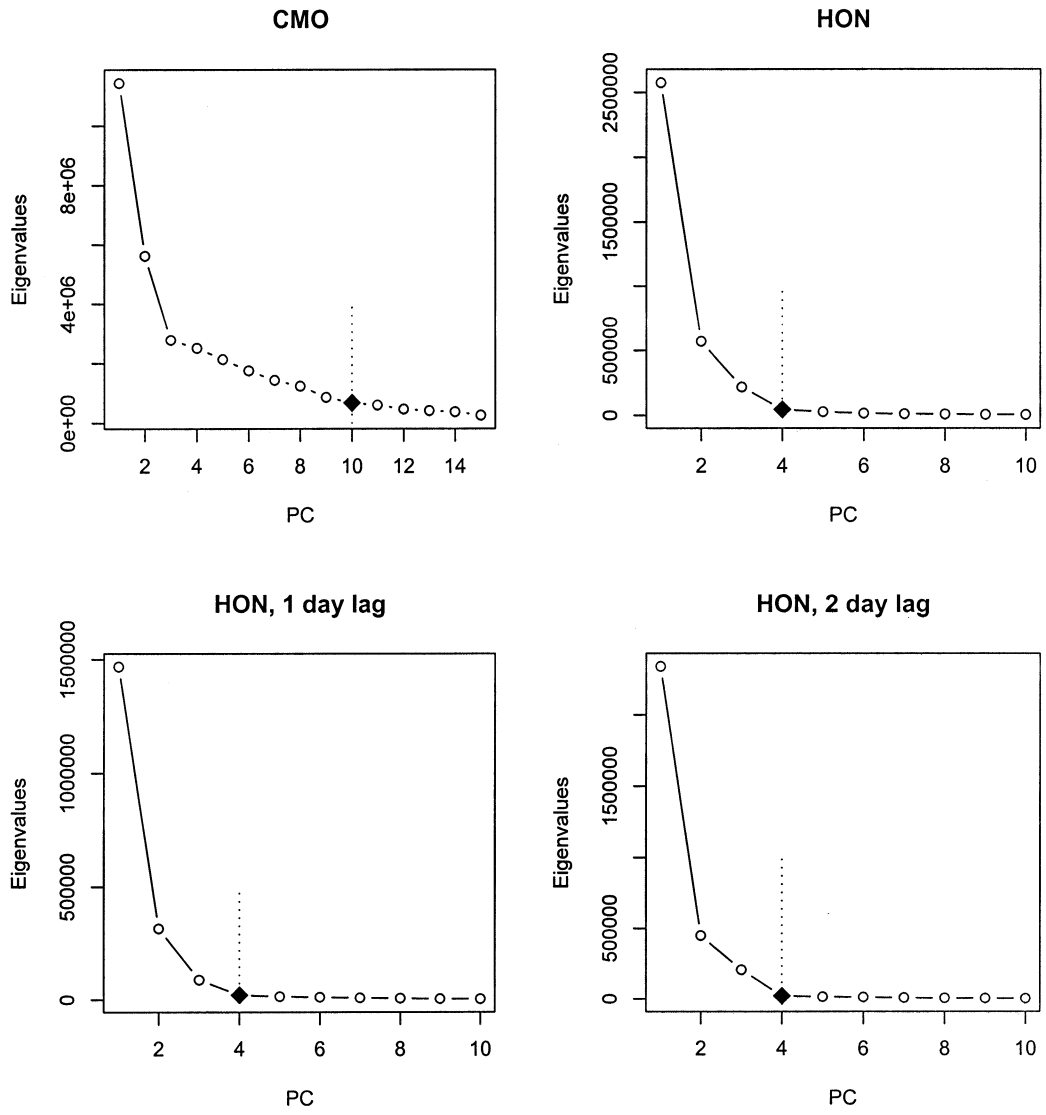


Fig. 2.6: Eigenvalues for different principal components of the substorm days that occurred from March until May, 2001, from College(CMO), Honolulu (HON) stations.

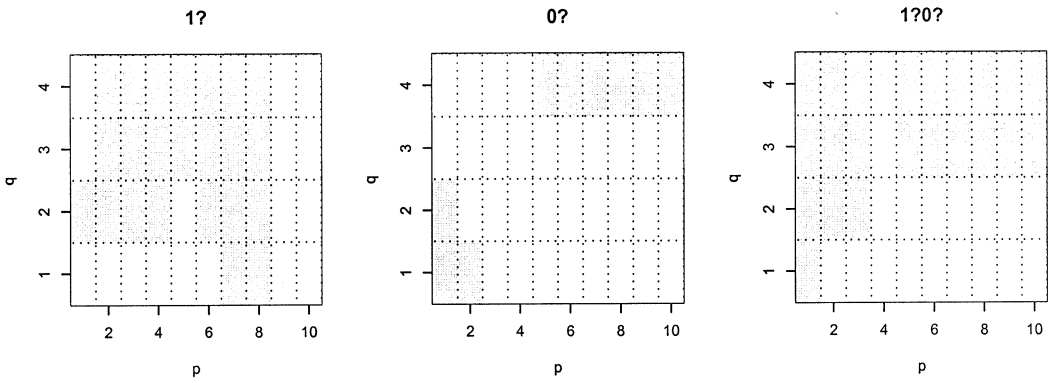


Fig. 2.7: Examples of rejection/acceptance plots at 5% level which are difficult to interpret. Grey area – reject H_0, white – fail to reject H_0.

CHAPTER 3

STATISTICAL SIGNIFICANCE TESTING FOR THE ASSOCIATION OF MAGNETOMETER RECORDS AT HIGH–, MID–, AND LOW–LATITUDES DURING SUBSTORM DAYS[1]

3.1 Introduction

Currents flowing in the magnetosphere-ionosphere (M-I) form a complex multiscale system in which a number of individual current components connect and influence each other (see [25, 26, 31]. The variabilities of these currents are closely connected to various nonlinear dynamic M-I processes, such as magnetic storms and substorms. Among the various observational means, the global network of ground-based magnetometers stands out with unique strengths [32]. About a hundred terrestrial geomagnetic observatories form a network, INTERMAGNET, designed to monitor the variations of the M-I current system. Modern digital magnetometers record three components of the magnetic field in five second resolution, but the INTERMAGNET's data we use consist of one minute averages, i.e. 1440 data points per day per component per observatory. We work with the Horizontal (H) component of the magnetic field (Chapter 13 of [2]), which reflects the variation of the M-I currents we plan to study. Figure 3.1 shows examples of magnetometer records we work with.[1]

Since the individual currents in the M-I system are connected and influenced by each other, enormous research efforts have been put into the study of the dynamical connection between the currents at high latitudes and low- mid-latitudes (e.g, [33]) as well as the storm-substorm relationship (e.g, [25]). In these endeavors, ground-based magnetometer data have been widely used to study how the high-latitude electrody-

[1]Coauthored by I. Maslova, P. Kokoszka, J.J. Sojka, and L. Zhu.

magnetometer data have been widely used to study how the high-latitude electrodynamics are manifested in the low- mid-latitude magnetic disturbances (e.g, [34]). It has been indicated by (see e.g. [17]) that the auroral currents may have, a perhaps indirect, impact on the equatorial and mid-latitude currents. The present paper focuses on the effects of the substorm current system at high latitudes on the magnetic disturbances at low- mid-latitudes, which has drawn wide interests over the years in the space science community, but with a new mathematical approach of the test of significance proposed in [35]. The approach we advocate directly uses the raw data curves of the H component, rather than derived indices. One of the shortcomings of the correlation analysis based on indices like Kp or Dst is that the physical interpretation of the indices is not obvious, see [36]. Moreover, while correlation analysis is useful as an exploratory tool, and it motivated our research, it does not allow to attach statistical significance to the conclusions if the data are dependent and non-Gaussian (like the H–component series).

The approach we propose here is novel in several ways: 1) we work directly with the measurements of the magnetic field, rather than indices; 2) we view magnetometer records over one day as a single *functional* observation; 3) we use a statistical test of significance, which by its very nature takes into account random variability not attributable to physical effects.

Viewing a whole daily curve consisting of 1440 data points as a single observation focuses the study on the interaction between the shapes of these curves. Typical shapes of storms or substorms are what has long been intuitively used; here we make an attempt to quantify this intuition by using a tool from the rapidly growing field of statistics known as *functional data analysis* (FDA), see [3], [4] for a comprehensive introduction, and [18] and [37] for related applications to magnetometer data. We hope that the method we propose will become useful in other geophysical studies

requiring statistical analysis of curves.

We test if equatorial or mid-latitude magnetometer records are statistically independent of the high-latitude records. This is our null hypothesis. If it is rejected, the test points toward a dependence of the low- or mid-latitude curves on the high-latitude curves. Thus, the acceptance of the null hypothesis is interpreted as the lack of effect of the auroral currents on the currents measured at low- or high-latitude. The rejection of the null hypothesis points towards the existence of a statistically significant association.

This paper shows that substorms have statistically significant influence on the currents measured at mid- and low-latitude. This dependence lasts for up to two days. After that period, the null hypothesis of zero effect generally cannot be rejected. We cannot specify a mechanism responsible for this surprising conclusion, and hope that our statistical analysis will stimulate further work.

The paper is organized as follows. In Section 3.2, we introduce the test of zero effect mentioned above. This is followed in Section 3.3 by a detailed analysis of magnetometer data. We start with the description of the data sets, and the motivation behind their selection is also provided. Then, technical details of the application of the test and the interpretation of the results are presented, followed by tabulation and discussion of the results. Main conclusions are summarized in Section 3.4.

3.2 Statistical Test of No Effect

In order to formulate a statistical test of significance, data must be assumed to follow a statistical model in which relationships between the data are specified up to some random errors. These errors contain, often unknown, effects which we do not wish, or are unable, to include in the model. One of the most popular FDA models is the functional linear model, see Chapters 12–17 of [4]. In our context it is appropriate

to work with the so called fully functional model defined by the equations

$$(3.1) \qquad Y_n(t) = \int_0^T \psi(t,s)X_n(s)ds + \varepsilon_n(t), \quad n = 1, 2, \ldots, N, \quad t \in [0, T].$$

The curves $Y_n(t)$ are the response variables, the curves $X_n(s)$ are the explanatory variables, and the $\varepsilon_n(t)$ are the error curves. Here, T is the length of one functional observation. In our case $T = 1440$ which is the length of one day in minutes. The function $\psi(t,s)$ is a kernel of an integral operator. This model is seen to be an extension of a simple linear regression to the case when the observations are curves rather than scalars, and can be succinctly written as

$$(3.2) \qquad Y_n = \Psi X_n + \varepsilon_n, \quad n = 1, 2, \ldots, N,$$

where Ψ is an integral operator in the Hilbert space of square integrable functions.

Assuming model (3.2), we wish to test

$$H_0 : \ \Psi = 0 \quad \text{versus} \quad H_A : \ \Psi \neq 0.$$

Using the analogy with a simple linear regression, we test if the slope of the regression line is zero, which means that the Y_n are independent of the X_n. For scalars, H_0 implies that the correlation coefficient is zero. In our case, the Y_n and the X_n are curves, and the problem is more complex. The null hypothesis, H_0, means that there is no linear association between the curves X_n and the curves Y_n; the alternative, H_A, that there is such association. In this paper, X_n are the high-latitude daily curves, and Y_n are daily curves from mid and low latitudes. The Y_n can be recorded during the same UT day, or subsequent UT days.

The testing procedure we use was developed by [35]. The test statistic $\hat{T}_N(p,q)$ is defined by

$$\hat{T}_N(p,q) = N \sum_{k=1}^{p} \sum_{j=1}^{q} \hat{\gamma}_k^{-1} \hat{\lambda}_j^{-1} \langle \Delta_N \hat{v}_k, \hat{u}_j \rangle^2 . \tag{3.3}$$

It depends on the number p of the most important principal components (PC's) of the curves X_n and the number q of the most important PC's of the Y_n, see below. The quantities $\hat{\gamma}_k$ and $\hat{\lambda}_j$ are the eigenvalues, and $\hat{v}_k$ and $\hat{u}_j$ are the PC's of the X_n and Y_n, respectively. The operator Δ_N is defined by $\Delta_N x = N^{-1} \sum_{n=1}^{N} \langle X_n, x \rangle Y_n$, where $\langle x, y \rangle = \int_0^T x(t)y(t)dt$ is the Hilbert space inner product. All the quantities appearing in (3.3) can be readily computed using the statistical software package R.

Theorem 3.1 of [35] shows that, as the samples size (number of pairs of curves) N increases, the distribution of $\hat{T}_N(p,q)$ converges to the well–known chi–squared distribution with pq degrees of freedom. The null hypothesis is rejected if $\hat{T}_N(p,q) > \chi^2_{pq}(\alpha)$, where $\chi^2_{pq}(\alpha)$ is the αth upper percentile of the χ^2_{pq} distribution. We use the standard $\alpha = 5\%$ significance level. Simulations in [35] show that the test is applicable for $N \geq 40$.

We conclude this section by explaining briefly the idea of principal component analysis (PCA), Chapter 8 of [4] contains a detailed exposition. PCA is an orthogonal linear transform that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first principal component, the second greatest variance – on the second, and so on. Principal components form an orthonormal system. The main goal of the PCA is to find the dominant modes of variation in the data. We want to know how many of these modes of variation are required to satisfactorily summarize the original data. Retaining only the characteristics of the data that contribute most to its variance will improve the

signal-to-noise ratio of what we keep. In Section 3.3.2, the p and q selection procedures are discussed.

3.3 Analysis of Magnetometer Data

3.3.1 Data description

We analyze one-minute records of the horizontal intensity of the magnetic field from four sets of stations given in Table 5.1. Each set consists of geomagnetic observatories that are roughly aligned along the same longitude, and are at different latitudes. The functional data consists of daily curves in UT time, with 1440 observations per curve. Figure 3.1 provides examples of such curves.

The question is whether the auroral activity reflected in the high-latitude curves affects the processes in the equatorial regions, reflected by mid- and low-latitude curves. To answer it, we analyze the effect of substorms recorded at College (CMO, latitude – 64.87, longitude – -147.86) on the records at other latitudes and longitudes.

Several types of data sets are analyzed. The first set consists of the days with substorms from January until August, 2001 (set A). Then the same analysis is performed on medium strength substorms (400-700 nT) during the same period (set B). In order to eliminate all possible storm effects in the data, from the substorm list we remove all days that contained storms, as well as, the days before and after it (set A*). We also removed such days from the list of medium strength substorms (set B*). We also considered only isolated substorm days, ie. substorm days followed by at least two quiet days (set I). Finally, we select the substorms that took place during three months, i.e. January – March (set C_1), March – May (set C_2), and June – August (set C_3), 2001.

The reason for performing the analysis on medium strength substorms separately

is that very strong substorms might distort the overall pattern of dependence. Removing days with storms additionally validates our findings by answering a possible criticism that what is seen is the effect of storms on substorms, and not the effect of substorms on low-latitude currents. As will be seen in the following, the most surprising finding of this work is that the effect of substorms lasts for up to two days. As no mechanism for such an dependence can be offered at this point, the conclusions may be criticized by saying that what we see one day after a substorm is the effect of next day's substorm. For this reason we found 33 isolated substorms, i.e. there are no substorms in the following two days. An example of an isolated substorm day (quantified by four auroral envelopes) is shown in Figure 3.2 together with two quiet days that follow it. Such extra precautions are not usual in statistical lagged correlation analysis, but are added here as an additional check. For example, if substorms occured on days 3, 4, 5, 10, 15, to study their effect 48 hours later, we compare them with low-latitude records, respectively, on days 5, 6, 7, 12, 17. The design of the test ensures that the substorms on days 4 and 5 are not compared to low-latitude records on day 5, only the substorm on day 3 is. If that were not the case, and substroms on consecutive days were similar, we would see dependence well beyond 24 hours, and we do not see it. We must also keep in mind that what we claim is the *average* statistical effect, and the conclusions do not apply to every single substorm. Finally, comparing the substorms over three-month periods ensures that the observations are approximately identically distributed, which is one of the assumptions of the test. The locations of the stations relative to the Sun change with season, so the substorms that happened long time apart might follow different statistical distributions. This is an additional precaution to validate our conclusions.

There are a couple of criteria for the selection of substorm days. First, we analyzed the features and the magnitude of H variations from College station. We

also used D and Z components, as well as, the AE index as a reference. The selection of substorm events was based on the following considerations. We included the substorm events such that a substorm was recorded at CMO. The days that were considered to be quiet were quiet in the global sense, because we also made sure that AE index did not show any disturbances. This fact was used in the study of isolated substorms described above. In the latter, we chose only the days when a substorm was recorded at CMO and there was quiet period of two days after that (no substorms anywhere according to the AE index). This procedure prevented the possibility of misclassifying the substorm days. Using the profile and character of the H variation as well as D and Z variations at CMO, and the AE index helped exclude non-substorm disturbances, such as pseudo substorms, as well as the substorms not seen at CMO. If there was no substorm recorded at CMO, but AE index showed a substorm, that day was excluded from the study. The classification of substorm was mainly based on the variation of H components at CMO stations, i.e strong – more than 700 nT, medium – 400-700 nT, weak – 200-400 nT.

There were 101 substorm days from January until August during 2001, 81 substorm of which did not have any storms around; 41 substorms were medium strength, 35 medium strength substorms after removing the ones close to the storms; 43 isolated substorms occurred during 2001. We observed 40 substorm days from January until March, 42 – from March until May, and 42 – from June until August. Note that here we use overlapping three-month periods, therefore the total number of substorms during those periods of time does not add up to the total number of substorms that took place from January until August. As indicated in Section 3.2, the test gives reasonable results with sample size starting from $N = 40$. Therefore, the samples used in our study are sufficient.

In order to perform the test, the minute-by-minute data were converted into

functional objects in R using splines basis with 149 basis functions. The number of basis functions is not crucial, the only requirement being that the smoothed curves should look almost identical to the original, while some noise can be eliminated. For more details on this procedure see Chapter 3 of [4].

3.3.2 Details of test application and interpretation

We now present some technical details needed to apply the test and properly interpret its outcomes.

In order to ensure that the test gives reliable results, the approximate validity of the functional linear model must be checked. For this purpose, a technique introduced by [12], which relies on a visual examination of scatter plots of scores, can be used. If the model is valid, score plots are roughly football-shaped. When the dependence is not linear, these plots exhibit different patterns. The number of plots is pq, where p and q are as in Section 3.2. They show the interaction of the kth PC of the X_n ($k = 1, \ldots, p$) and jth PC of the Y_n ($j = 1, \ldots, q$). To illustrate this technique, consider a non-linear model: $Y_n(t) = H_2(X_n(t)) + \varepsilon_n(t)$, where $H_2(x) = x^2 - 1$ is the Hermite polynomial of rank 2. For this model, the plot in the top left corner of Figure 3.3 exhibits a quadratic trend. For model (3.2) to be valid all these plots should be "pattern-free", i.e. football-shaped. Figure 3.4 shows examples of these plots for magnetometer data. We used CMO medium strength substorm records as X, and THY with no lag – as Y. These scatter plots indicate linear relationship with some outliers. Since we do not require Gaussianity, only finite fourth moment, these outliers need not invalidate our conclusions. In case of other pairs of functional data, the score plots look similar. We conclude that a linear model is approximately appropriate for our application.

We now describe how to choose the most important PC's that will be used

in the test. One of the ways to pick the most important PC's is to use the scree test, which is a graphical method first proposed by [30]. To apply the scree method one plots the successive eigenvalues against the corresponding PC (see Figure 3.5). The method suggests to find the place where the smooth decrease of eigenvalues appears to level off to the right of the plot. To the right of this point one finds only "factorial scree" ("scree" is a geological term referring to the debris which collects on the lower part of a rocky slope). Table 3.2 provides the number of most important principal components and corresponding percentage of total variability explained by them during substorms that occurred from January until August, 2001. For other data sets under consideration the general pattern is similar and it is available upon request. One can also notice from Figure 3.6 that each subsequent component picks up variation that declines in smoothness. For example, the 10th principal components resemble random noise and explain a small percentage of variability, that is why they are not included in the analysis.

When testing the no effect hypothesis, in most cases there is a clear rejection or acceptance for all combinations of the most important principal components. In those cases, we can either reject "1" or fail to reject "0" the null hypothesis with a reasonable confidence. In this paper we use the nominal 95% confidence level. However, there are some cases when it is not clear what conclusion to draw. We denote such cases "1?" – inclined toward rejecting the null hypothesis, "0?"– inclined toward failing to reject the null, "1?0?"– inconclusive. Figure 3.7 gives examples of such cases. We plot rejection regions up to the number of important principal components. Grey areas mean that we reject H_0, white – fail to reject H_0. The conclusion is clear when all, or almost all, rectangles are of the same color. We can then conclude that X has an effect on Y (all grey) or there is no effect (all white). Left panel of Figure 3.7 gives an example when it is not clear what to conclude. However, based on our previous

experience we are most likely to reject the null hypothesis. In the case shown in the middle panel, the conclusion is also not clear, but we lean toward accepting the independence of X and Y. Finally, the right panel presents an example where it is unclear what to conclude. All the methods introduced above led to the results presented in the following section.

3.3.3 Testing for substorm effect

We now discuss the results of the application of the test. We consider high-latitude records from College station (CMO) as X, and let Y be the observations from all eight mid- and low-latitude stations during the same UT time as the CMO data. We also analyze responses one, two, and three days after substorms were recorded at the CMO station. The notation we use later indicates the station code and the number of lagged days, e.g. BOU0 – lag 0; BOU1 – lag 1; BOU2 – lag 2; BOU3 – lag 3. Such a setting should allow us to see if there is a longitudinal effect of substorms; how long this effect, if any, lasts; and what the global influence of a substorm is.

Column A in Table 3.3 presents the test results for all the substorms that occurred from January to August. We see that the effect of substorms observed at CMO is statistically significant at all mid- and low-latitude stations at the same UT (e.g. BOU0, HON0). This is true for one-day lag data as well (e.g. BOU1, HON1), but for the lag of two days the results are inconclusive. We conclude that the effect of substorms observed at CMO persists for about 48 hours, at all longitudes and latitudes. In the column labeled A* we provide the test results for the set of the substorms where none of the events occurred close to storms. (Storms were identified based on the shape and duration of the bay variation below -100nT, in some cases we checked if there was a sudden commencement.) As one can see, the results are similar

to the ones in column A. This means that the observed association is not attributable to an impact of storms on high-latitude currents. We also analyzed the effects of isolated substorms, i.e. there were at least two quiet days after such substorms (see column I in Table 3.3). As one can see, there is significant linear dependence between records observed at high latitude and mid-, low-latitude during substorm days, as well as the next day. This means that the next day association cannot be attributed to the confounding effect of substorms on consecutive days. Next, we analyze the effect of medium strength substorms. Table 3.3, column B, presents the test results. We can see that the medium strength substorm effect is weaker than in case of all substorms. The effect of medium strength substorms appears significant on the same day, but on the following days is absent. It fades out faster for further longitudes. We draw the same conclusion from column B* which includes test results on the medium strength substorms that were not effected by the storm activity. Table 3.4 gives the results for the three sets of substorms in three-month periods. In column C_1 the results for the substorm days from January to March, 2001 are presented. The conclusions are similar to the ones we got for all substorms from January until August (see Table 3.3, column A). The dependence seems to last for two days, i.e. the day when the substorm occurred and the following day. We come to the same conclusion dealing with the other two sets of the substorms, the ones that occurred in Spring and Summer 2001 (see columns C_2 and C_3 of Table 3.4), the second day dependence being weaker in summer. This agrees with the earlier analysis, as there are fewer strong substorms in summer months.

We conclude that there is a pattern that suggests that there is a dependence between high- and mid-, and high- and low-latitude records with no and one day lag. There is no significant dependence for data with two- and three-day lags.

The form of the integral kernel $\psi(s,t)$ in (3.1) is quite complex. It cannot therefore be hoped that the shape of the X_i will in some way be reproduced in the shape of the responses Y_i. Figure 3.8 shows examples of the surfaces $\psi(s,t)$ estimated using spline expansions, see [38] for the details.

3.4 Conclusions

This paper provides a novel analysis of the impact of the auroral currents on the currents at lower latitudes. Our technique is akin to the ideas of [18]. It indicates that there is significant association between substorms recorded at a high-latitude station (CMO) and the magnetic records at mid- and low-latitudes and all longitudes. This dependence disappears two days after the substorms and is weaker the further we get from the CMO station. It decreases faster for moderate substorms, for which only the same day effect can be claimed.

Some discussion of these findings is in order. The ground magnetic effects of a localized auroral current system in the ionosphere normally become insignificant for a location at the Earth's surface 400-500 km away from the center of the current system. Therefore the substorm auroral currents in the ionosphere would not be expected to have significant *direct* effects on the measurements of the B field at mid-latitudes and most certainly not at low (equatorial) latitudes. However, given that the analysis finds that indeed the most important principal components of the high- and mid/low-latitude magnetic field are correlated, it implies that the influence is not directly from the auroral electrojets, but the full current circuit in the M-I system that drives the auroral electrojets during substorms. Conceptually, this would not be entirely unexpected. However, what is unexpected is that on the subsequent day, after a 24-hour lag, the mid- and low-latitude field is still correlated with prior day's substorm activity defined by high-latitude magnetic fields. The result is dependent

on the strength of the substorms, i.e. only the effect of strong substorms extends to low latitudes on the second day. The interpretation of this result is not readily apparent. That a current system has a 24-hour lagged memory seems at odds with substorm electrodynamics, or that substorm energy disposition at high latitudes can actually be propagated to mid and low-latitude to modify currents in the ionosphere (Sq type or dynamo type) with a 24-hour lag is also unlikely. These statistical findings may imply some physical connections between the substorm electrodynamics and the physical processes in other regions of the M-I system that we are not aware of at the present time. The stage is set for follow-on analysis to provide more extensive insight into the nature of these dependencies.

Table 3.1: Geomagnetic observatories used in this study.

Latitude	I	II	III	IV
High	College (CMO)	–	–	–
(Lat, Lon)	(64.87, -147.86)	–	–	–
Mid	Boulder (BOU)	Fredericksburg (FRD)	Tihany (THY)	Memambetsu (MMB)
(Lat, Lon)	(40.14, -105.24)	(38.20, -77.37)	(46.9, 17.89)	(43.90, 144.20)
Low	Honolulu (HON)	San Juan (SJG)	Hermanus (HER)	Kakioka (KAK)
(Lat, Lon)	(21.32, -158.00)	(18.11, -66.15)	(-34.43, 19.23)	(36.23, 140.18)

Table 3.2: Number of principal components retained by the scree test, and percentage of total variability explained, during substorm days that occurred from January until August, 2001.

Stations	PC	%	Stations	PC	%	Stations	PC	%	Stations	PC	%
CMO	10	82.52									
BOU0	5	91.36	FRD0	4	90.83	THY0	5	92.17	MMB0	4	92.30
BOU1	4	86.40	FRD1	4	89.55	THY1	5	89.49	MMB1	4	91.01
BOU2	4	91.17	FRD2	4	92.32	THY2	4	91.57	MMB2	4	94.59
BOU3	4	91.74	FRD3	4	92.68	THY3	4	91.51	MMB3	4	95.60
HON0	4	96.56	SJG0	5	97.08	HER0	4	95.07	KAK0	4	94.33
HON1	3	94.91	SJG1	4	94.57	HER1	4	94.31	KAK1	4	93.80
HON2	4	97.44	SJG2	3	92.73	HER2	4	95.89	KAK2	4	96.39
HON3	4	97.79	SJG3	4	96.42	HER3	4	95.53	KAK3	3	94.66

Table 3.3: Results of the test for all substorm days (A), substorm days excluding days around the day with a storm (A*); medium strength substorms (B), medium strength substorms excluding storm days (B*) that occurred from January to August, 2001; (I) isolated substorms that occurred from January to December, 2001.

Mid-latitude																			
A	A*	I	B	B*	A	A*	I	B	B*	A	A*	I	B	B*	A	A*	I	B	B*
		BOU0					BOU1					BOU2					BOU3		
1	1	1	1?	1?	1	1	1	0	1?	1?0?	1?0?	0	0	0	0	1?0?	0	0?	1?0?
		FRD0					FRD1					FRD2					FRD3		
1	1	1	1?	1?0?	1	1	1	0	0?	0?	0?	0	0	0	0	0?	0	0?	1?0?
		THY0					THY1					THY2					THY3		
1	1	1	1?	1?	1	1	1	0?	1?0?	1?0?	1?0?	0	0	0	0	0?	0	0?	1?0?
		MMB0					MMB1					MMB2					MMB3		
1	1	1	0?	1?	1	1	1	0	1?0?	1?	1?	0?	0	0	0?	0	0	0?	1?0?
Low-latitude																			
A	A*	I	B	B*	A	A*	I	B	B*	A	A*	I	B	B*	A	A*	I	B	B*
		HON0					HON1					HON2					HON3		
1	1	1	1?	1?	1	1	1	1?0?	1?	1?0?	0?	0	0	0	0?	0?	0	0?	1?0?
		SJG0					SJG1					SJG2					SJG3		
1	1	1	1?	1?	1	1	1	0	0?	0?	0	0	0	0	0	0	0	0?	1?0?
		HER0					HER1					HER2					HER3		
1	1	1	0?	1?	1	1	1	0?	0?	1?0?	1?0?	0	0	0?	0	0?	0	0?	1?0?
		KAK0					KAK1					KAK2					KAK3		
1	1	1	1?	1?	1	1	1	1?	1?	1?	1?	1?0?	0	0	0	0?	0	0?	1?0?

Table 3.4: Results of the test for substorm days that occurred in 2001 from January to March (C_1), March to May (C_2), June to August (C_3).

Mid-latitude											
C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
BOU0			BOU1			BOU2			BOU3		
1	1	1	1	1	1?0?	0	0	0	0	0	0
FRD0			FRD1			FRD2			FRD3		
1	1	1	1	1	1?0?	0	0	0	0	0	0
THY0			THY1			THY2			THY3		
1	1	1	1	1	1	0	0	0	0	0	0
MMB0			MMB1			MMB2			MMB3		
1	1	1	1	1	1?0?	1?0?	0	0?	0?	0	0
Low-latitude											
C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
HON0			HON1			HON2			HON3		
1	1	1	1	1	1?0?	0?	0	0	0	0	0
SJG0			SJG1			SJG2			SJG3		
1	1	1	1	1	1?	0	0	0	0	0	0
HER0			HER1			HER2			HER3		
1	1	1	1	1	1	0	0	0	0	0	0
KAK0			KAK1			KAK2			KAK3		
1	1	1	1	1	1	1?0?	0	0	0?	0	0

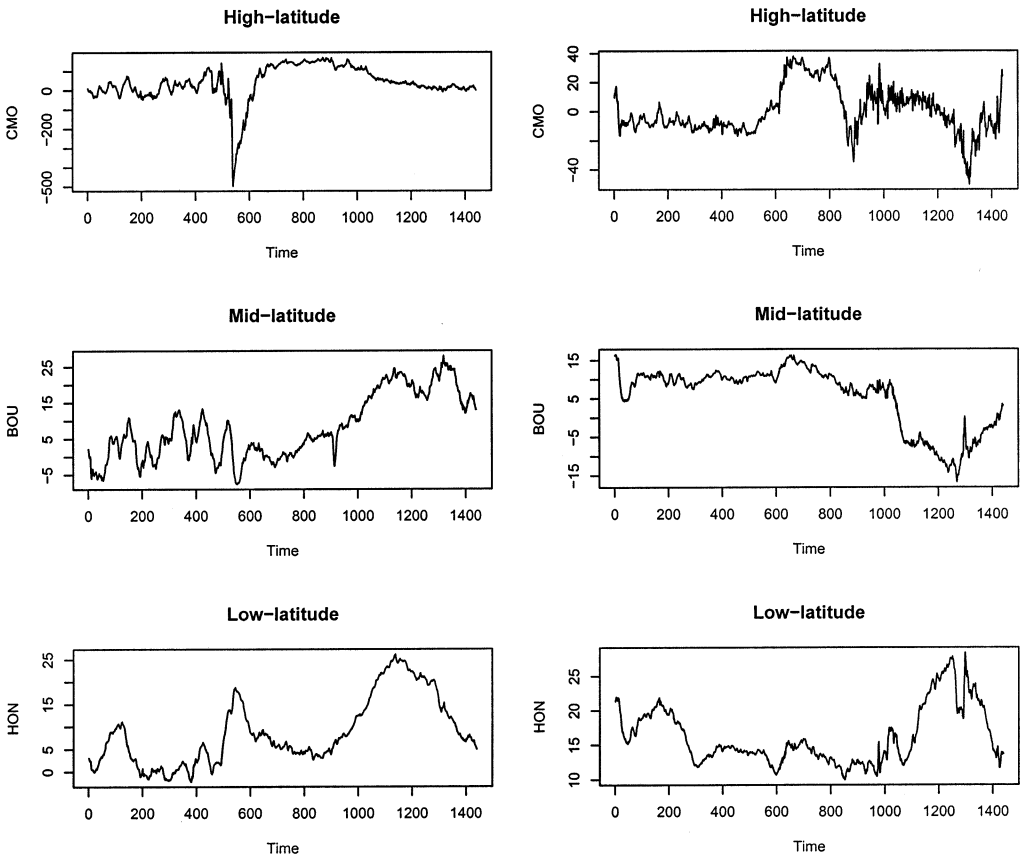


Fig. 3.1: Horizontal intensities of the magnetic field measured at a high-, mid- and low-latitude stations (College, Boulder, Honolulu) during a substorm (left column) and a quiet day (right column). Note the different vertical scales for high-latitude records. Each graph is a record over one day, which we view in this paper as a single *functional* observation.

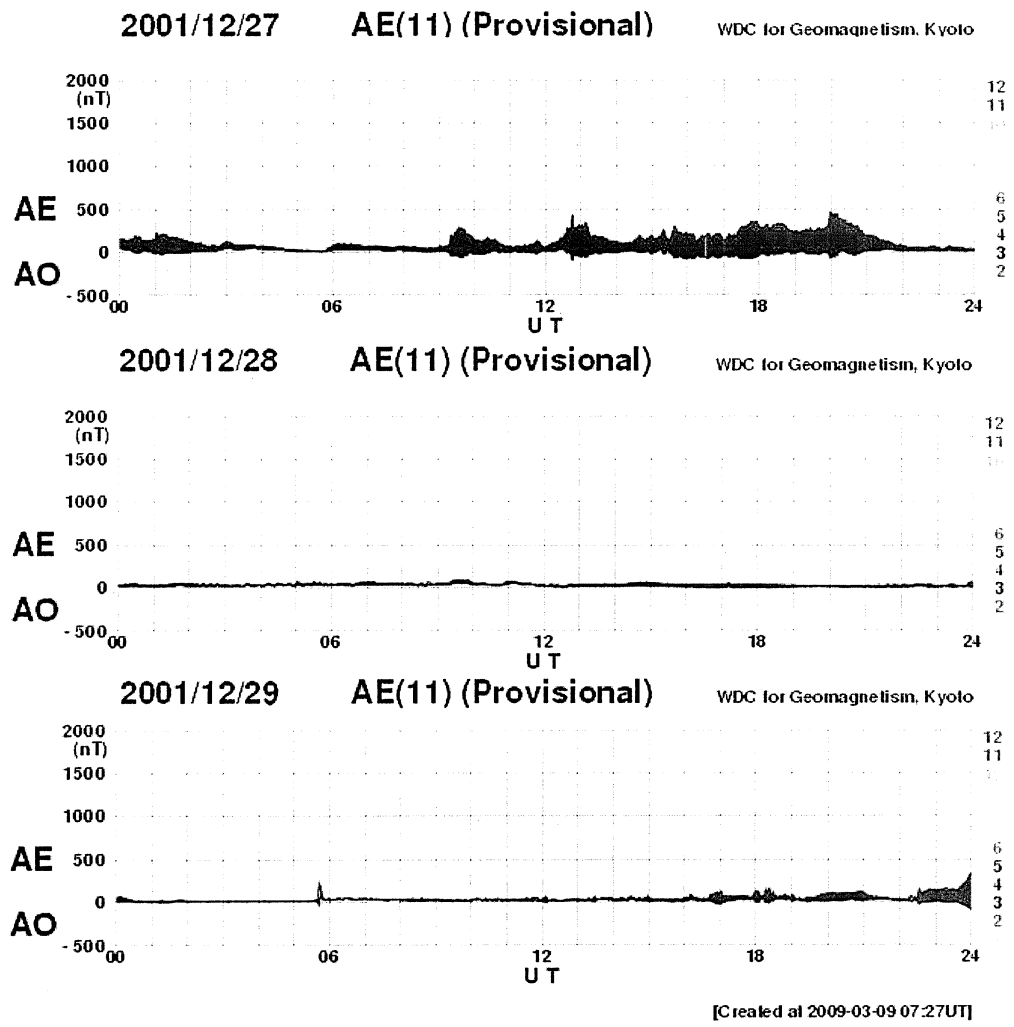


Fig. 3.2: AE index. An example of isolated substorm that took place on Dec 27 , 2001 (top panel) and two quiet days after it, Dec 28-29, 2001 (middle and bottom panels).

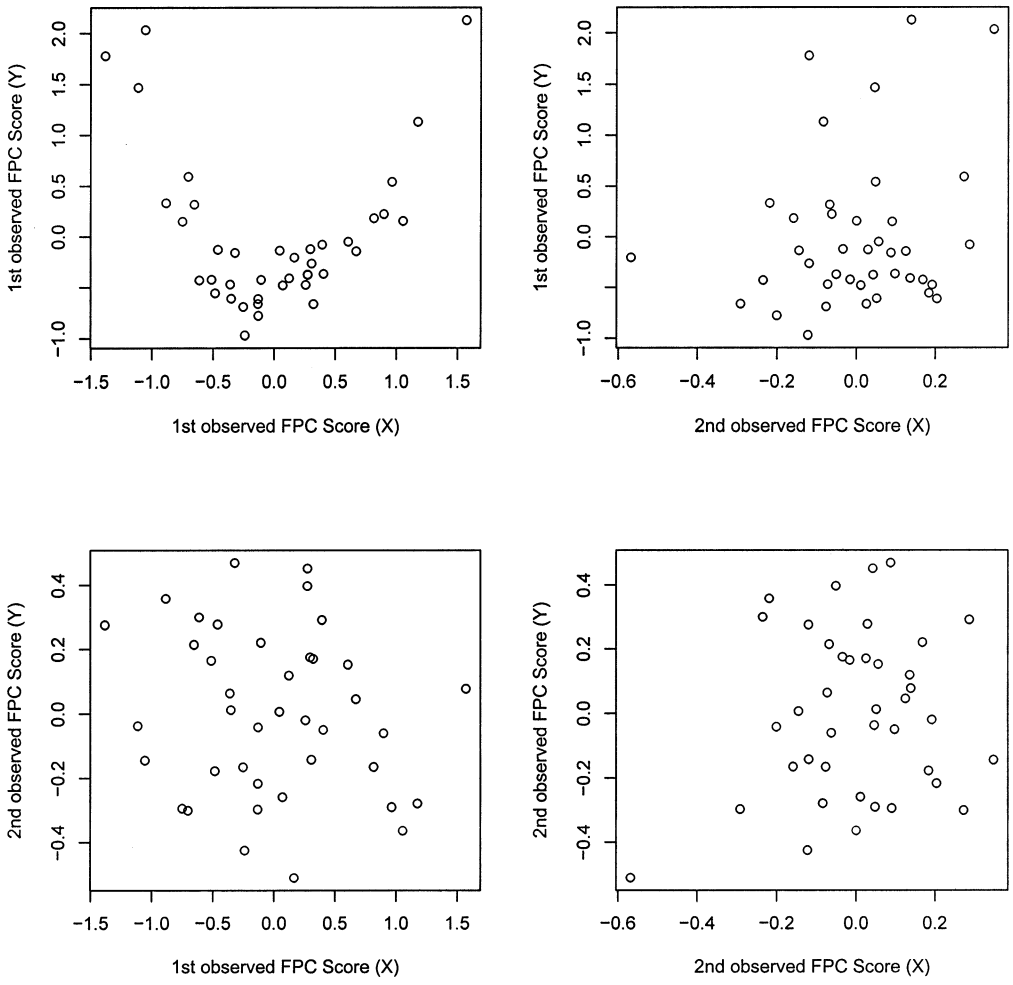


Fig. 3.3: Functional predictor-response plots of functional principal component scores of response functions versus functional principal component scores of predictor functions for $Y_n(t) = H_2(X_n(t)) + \varepsilon_n(t)$, where $H_2(x) = x^2 - 1$, $n = 1, \ldots, 40$.

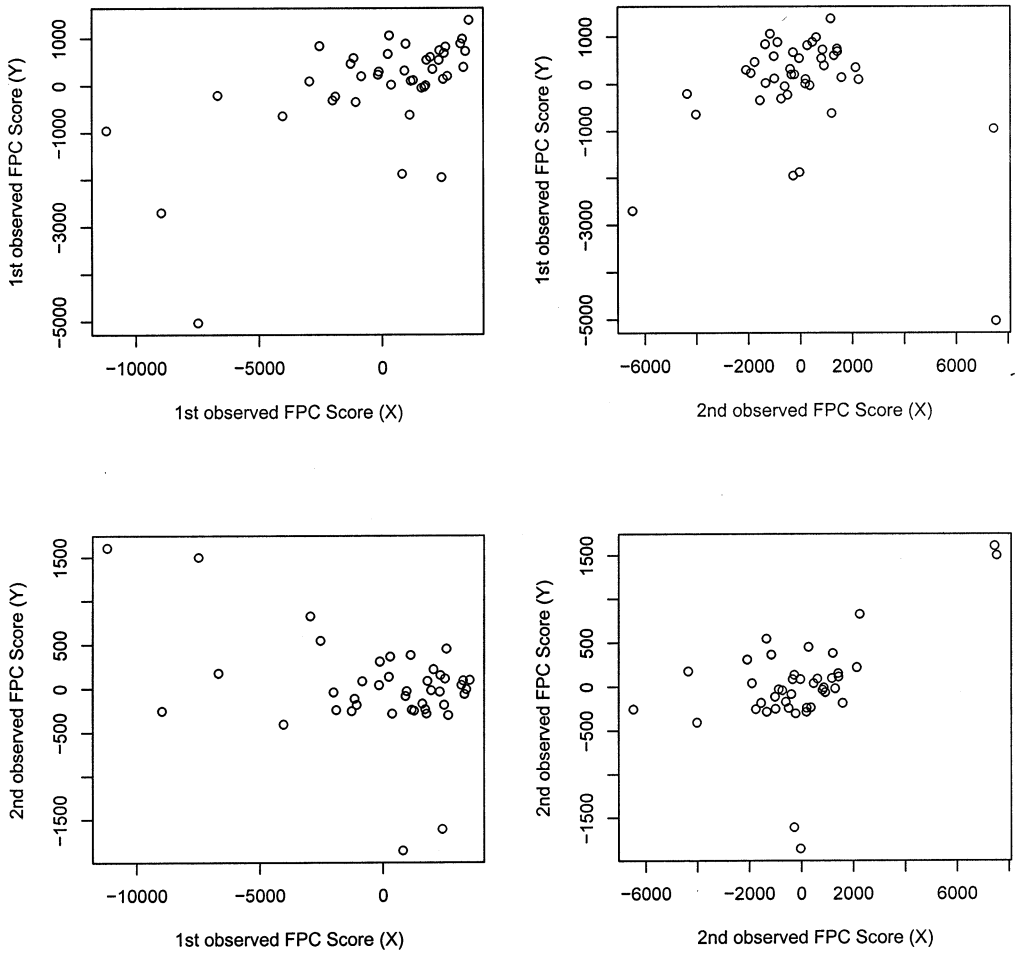


Fig. 3.4: Functional predictor-response plots of functional principal component scores of response functions versus functional principal component scores of predictor functions for magnetometer data (CMO vs THY0).

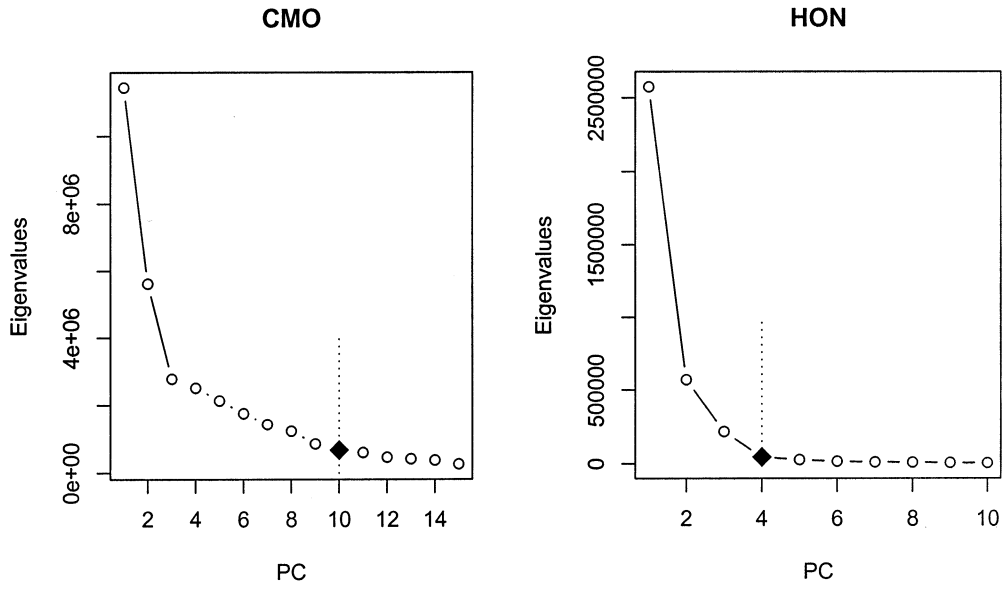


Fig. 3.5: Eigenvalues for different principal components of the substorm days that occurred from March until May, 2001, from College(CMO), Honolulu (HON) stations. The black diamond denotes the number of most important principal components selected by the scree test.

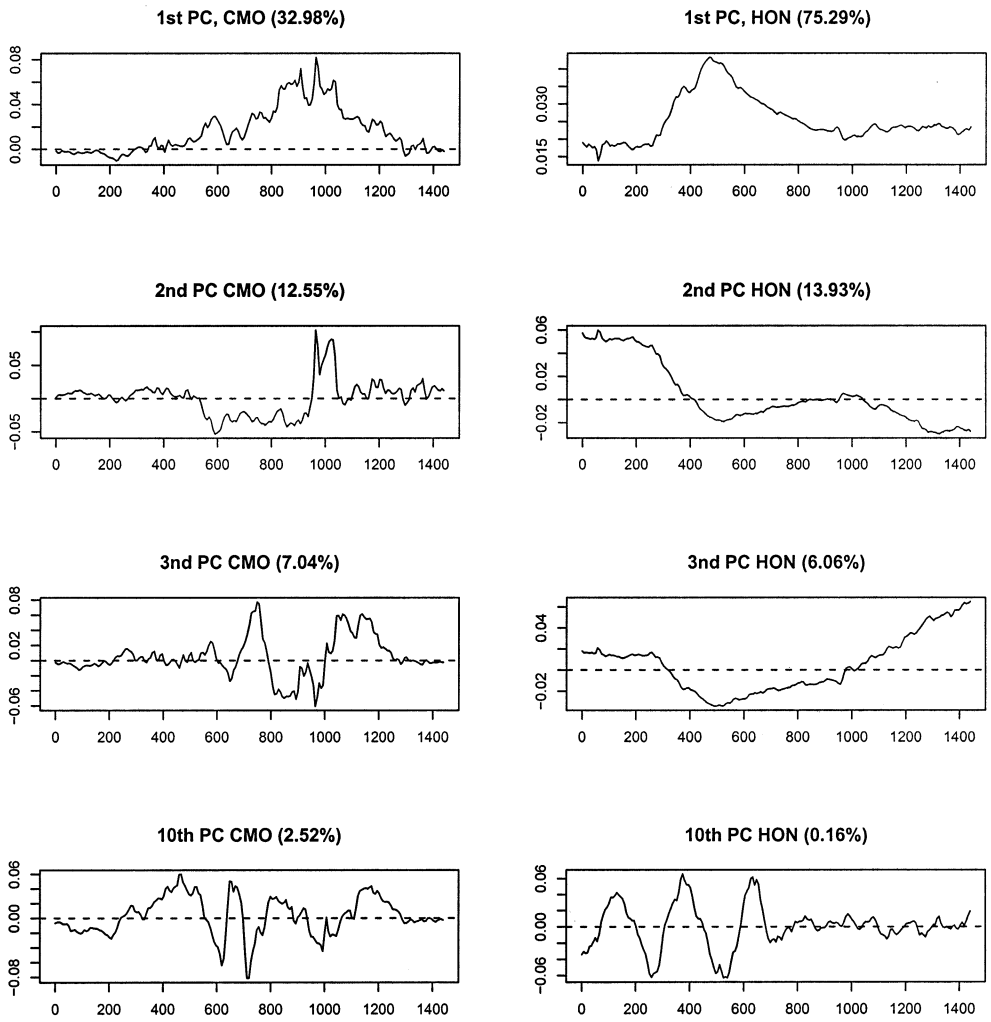


Fig. 3.6: Principal component curves (harmonics) of the substorm days that occurred from January until August, 2001, from College(CMO), Honolulu (HON) stations.

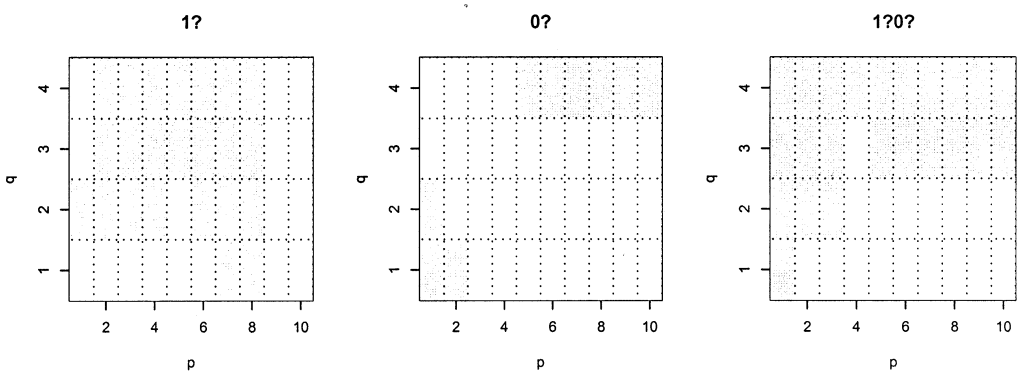
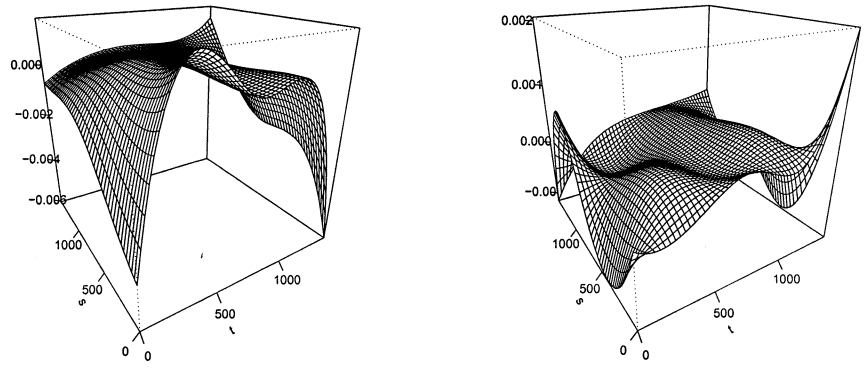


Fig. 3.7: Examples of rejection/acceptance plots at 5% level which are difficult to interpret. Grey area – reject H_0, white – fail to reject H_0.



(a) (b)

Fig. 3.8: Estimated surface $\psi(t, s)$. Here, $X_i(s)$ are the records from CMO station during days with an isolated substorm and $Y_i(t)$ curves are: (a) HON during the same time as CMO observations, (b) HON next day.

CHAPTER 4

REMOVAL OF NONCONSTANT DAILY VARIATION BY MEANS OF WAVELET AND FUNCTIONAL DATA ANALYSIS[1]

4.1 Introduction

It has long been recognized that even on quiet days the daily variation changes very visibly from day to day, both in its amplitude and its shape. This is attributable to multiple dynamic drivers which include not only tidal ionospheric winds, but also the effect of the Chapman-Ferraro current, the Sq current, and the magnetotail current, (see [18]) and references therein. On storm days, the effects of these drivers change even more and lead to a very complicated daily variation [39] which is difficult to deconvolute from the global effect of the intensified symmetric ring current. Yet, the Dst and related indices remove a *constant* daily variation from daily H-component signatures. In the Dst, this constant quiet variation is obtained by averaging the daily H-component curves over several quiet days in a month. The main goal of this paper is to introduce a technique allowing to remove daily variations which change from day to day, as well as the effects of other local-time dependent components, thus leading to an index of storm activity which better reflects the variability of the symmetric ring current.

Our approach builds on the WISA index introduced by [19] and studied by [40] and [41]. WISA is a one minute resolution version of the Dst index, and when appropriately averaged, is statistically indistinguishable from the one hour Kyoto Dst. It enjoys however important operational advantages over the Dst. It can be

[1]Coauthored by I. Maslova, P. Kokoszka, J.J. Sojka, and L. Zhu. Reproduced by permission of American Geophysical Union, Journal of Geophysical Research, Vol. 114, A03202, doi:10.1029/2008JA013685, 2009. Copyright [2009] American Geophysical Union.

over a year, and requires only the selection of equatorial terrestial observatories as input (no selection of quiet days is required). Nevertheless, the WISA procedure also removes a constant daily variation, which is just computed using a different, wavelet–based, algorithm. In order to construct a "cleaner" index of the storm activity, we need to remove the daily variations that are different each day. We propose a general, automatic technique that involves wavelet and principal component analysis methods and extracts a non-constant daily variation.

This work also builds on the ideas introduced by [18] who use the method of natural orthogonal (principal) components to analyze the daily magnetic variation, and argue that the first eigenmode represents the solar quiet daily variation. We address this matter in more detail in Section 4.3. [37] also use principal component techniques to separate Sq from complicated disturbances. However, as we show in the following section we believe that the procedure these authors use includes storm features in their estimated daily variation. In our paper we use similar techniques, but we argue that our proposed periodic component estimation methodology is more accurate.

The paper is organized as follows. A brief description of the requisite statistical concepts is provided in Section 4.2. In Section 4.3, we provide a detailed description of the construction of the improved index. Then, in Section 4.4, we compare the new index to WISA by means of functional canonical correlations. Finally, main conclusions are summarized in Section 4.5.

4.2 Wavelet and Functional Data Analysis

In this section we first introduce some basic ideas of the wavelet analysis focusing only on the aspects that are relevant to our task. Then we present the functional data analysis, mainly the functional principal component analysis.

First, we introduce a wavelet-based representation of the magnetometer data in order to explain the central ideas of the new procedure. Let, $X_s = \{X_{s,t}, t = 1, \cdots, N\}$ be the magnetogram recorded at station $s = 1, \cdots, m$, where N is the sample size recorded in minutes. We can write it as

$$X_s(t) = \sum_{j=1}^{J} D_{s,j}(t) + S_{s,j}(t),$$

where $D_{s,j} = \{D_{s,j}(1), \cdots, D_{s,j}(N)\}$ are the details, and $S_{s,j} = \{S_{s,j}(1), \cdots, S_{s,j}(N)\}$ is the smooth. Here, $j = 1, \cdots, J$ is the multiresolution analysis (MRA) level. The details capture the part of the records that correspond to the frequencies in the range from 2^{-j-1} to 2^{-j} cycles per minute. For further details see Chapter 5 of [42]. In this paper, we introduce a procedure that allows us to isolate the storm activity by applying statistical techniques to different levels j. We focus on detail levels $j = 8, 9, 10$. As explained in [19], these are the levels that contain daily periodic features: $j = 8$ captures approximately 6 hour periodic component, $j = 9$ — 12 hour component, and $j = 10$ — 24 hour component. We use the same transform and filter as used for WISA construction, i.e. the maximum overlap discrete wavelet transform (MODWT) and the LA(8) filter.

We continue this section by explaining briefly the idea of functional principal component analysis (FPCA), Chapter 8 of [4] contains a detailed exposition.

Panel (a) of Figure 4.1 shows an example of the data which is the sum of the MRA of magnetometer records at levels $j = 8, 9, 10$. Dashed lines indicate UT midnight. Our goal is to remove the daily component, so it is natural to split the data into daily observations (functions). The functional observations defined on 24 hour intervals are shown in panel (b) of Figure 4.1. Hence, we treat the daily records as functions and extract the daily variations using FPCA.

In multivariate case we define sets of normalized weights to emphasize types of variation that are most strongly represented in the data.

Let $u_m = (u_{1m}, \cdots, u_{pm})'$ be the m^{th} weight vector such that $f_{im}^{\cdot} = \sum_j u_{jm} x_{ij}$ has the largest mean square $\frac{1}{N}\sum_i f_{im}^2$ subject to constraint $||u_m||^2 = 1$, and $\sum_j u_{jk} u_{jm} = 0$, $k < m$, i.e. each mode must be orthogonal to the previous one so that they are indicating something new. We carry out the procedure, up to a limit of number of variables p.

The main idea of the FPCA is to find *functions* whose inner products with the data yield maximum variation in the curves. FPCA is an orthogonal linear transform that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first principal component, the second greatest variance – on the second, and so on.

In functional context $u(t)$ and $X(t)$ are functions and summation over j is replaced by integration over t. Similar to the multivariate case we define the j^{th} principal component score of X_i as $\gamma_j = \int X_i(t) u_j(t) dt$. It can be interpreted as the weight of the contribution of the FPC u_j to the curve X_i. Each principal component, say j^{th} $u_j(t)$ is chosen to maximize $\frac{1}{N}\sum_i (\int u_j(t) X_i(t))^2$ subject to constraint: $\int u_j^2(t) dt = 1$. Same as for multivariate PCA, the weight function $u_j(t)$ is required to satisfy the orthogonality constraint $\int u_k(t) u_j(t) dt = 0$, $k < j$.

Principal components form an orthonormal system. The main goal of the FPCA is to find the dominant modes of variation in the data. In this study we are interested in the first principal component. [18] and [37] argue that it captures the main features of the daily Sq variation. However, our approach differs from the one introduced by [18]. Instead of the raw data, we use filtered records, i.e. the sum of three levels of MRA, and remove storm features with care.

4.3 Removal of the Daily Variation

In this section we provide the details on the removal of the daily component from the magnetometer data. During quiet periods, it is basically the Sq variation, but during disturbed periods it may reflect the dynamo effect and disturbances from other M-I currents. This component is semi-periodic, as it is caused by the rotation of the Earth. The atmospheric dynamo generates currents that flow in the upper atmosphere in the E region. These currents arise as a consequence of atmosphere storm dynamics (winds) which have been generated by geomagnetic storms. During storms these wind patterns are quite different from the quiet time winds that create the Sq current system.

The periodicity is clearly visible in MRA details $D_{s,j}$ for levels $j = 8, 9, 10$ (see Figure 4.2). However, one can also see that it is enhanced during a storm. This fact is taken into account while removing the daily variation from storm features.

Let

$$D_{s,P}(t) = D_{s,8}(t) + D_{s,9}(t) + D_{s,10}(t), \quad t = 1, \cdots, N$$

be the part of the signal that includes practically all frequencies of the daily component spectrum. The subscript "P" stands for (semi) "periodic".

Our goal is to extract the signature of the storm activity from the $D_{s,P}$. A storm is a global event and it is visible in the records of all stations. Its signatures at various stations are aligned in UT. Figure 4.3 illustrates the fact that the storm features are roughly aligned in UT, whereas the periodic components are not. Therefore, we want to extract as many UT aligned features as possible. Such features are attributable to the storm activity, and should be included in the index.

As mentioned above, the periodic daily component is a local feature that is approximately aligned in LT. In order to separate it from the storm signature, we

first need to remove all features aligned in UT. Therefore, we remove

$$\overline{D}_P(t) = \frac{1}{m}\sum_{s=1}^{m} D_{s,P}(t), \quad t = 1, \cdots, N,$$

which is the average of $D_{s,P}$ of all stations $s = 1, 2, \cdots, m$ used in the study. The mean $\overline{D}_P$ roughly follows the storm pattern (see the thick line in Figure 4.3). After the mean removal, the data is mostly cleaned from events aligned in UT (see Figure 4.4) and so the daily variation can be removed more effectively. We emphasize, that the computation of $\overline{D}_P(t)$ is merely a preliminary step. As the records from different stations may have slightly different dynamic ranges, the average may be biased towards some stations. Averaging with appropriately computed weights is possible, but this increases the complexity of the algorithm, and leads to negligible gains.

Denote by $D^c_{s,P}(t) = D_{s,P}(t) - \bar{D}_P(t)$ the centered record at station s. Figures 4.5 (disturbed period of time) and 4.6 (quiet time) show that $D^c_{s,P}$ contains a strong quasi–periodic component which reflects the Sq variation during quiet periods and a more complicated Sq variations during storm periods. In Figure 4.5 we see that there is nighttime activity in $D^c_{s,P}$. We want to add it to the storm index, but not the quasi–periodic component. We therefore postulate that

$$D^c_{s,P}(t) = P_s(t) + R_s(t), \quad t = 1, \cdots, N, \tag{4.1}$$

where P_s is identified with the daily periodic component and R_s is the remaining effect of a storm left after the average removal. Next, we apply principal component analysis techniques to estimate the daily variation P_s. We convert $D^c_{s,P}$ into functional object, i.e. daily functions that start at UT midnight. Using principal component

analysis we can write (t' is the time in minutes within one day)

$$D^c_{s,P}(t') = \mu_s(t') + \sum_{j=1}^{\infty} \gamma_{s,j} u_{s,j}(t'), \quad t' = 1, \cdots, 1440,$$

where $\mu_s(t')$ is the daily mean, $\gamma_{s,j}$ is a score vector for j^{th} PC, and $u_{s,j}$ is the j^{th} PC for station s. We assume that periodic component for day $i = 1, \cdots, N/1440$ is

$$P_{s,i}(t') = \mu_s(t') + \gamma^*_{s,1,i} u_{s,1}(t'), \quad t' = 1, \cdots, 1440, \tag{4.2}$$

where $\gamma^*_{s,1,i}$ is a filtered score for the i^{th} day described below. The function $u_{s,1}(t')$ is the first PC for station s. In (5.3) $\mu_s(t')$ and $u_{s,1}(t')$ are deterministic functions defined over the 24-hour interval, and $\gamma^*_{s,1}$ are random weights that change from day to day. Hence, the extracted daily component $P_s(t)$ is non-constant. Note that $P_{s,i}(t')$ where $t' = 1, \cdots, 1440$ and $i = 1, \cdots, N/1440$ is the same daily periodic component as $P_s(t)$ where $t = 1, \cdots, N$ split into daily functions.

Decomposition (5.3) is akin to the ideas of [18] and [37], who argued that the first principal component follows the pattern of the daily Sq-variation. However, while these authors work with the raw magnetometer records, we first apply a wavelet filter to the data and use just the levels that contain the periodic component. We also remove the average of several stations to separate the storm effect. So, in our paper, to estimate daily periodic component P_s, we compute the first PC of $D^c_{s,P}$ rather than the first PC of the raw magnetometer data.

As mentioned earlier, the three selected MRA levels contain residual storm features. Averaging over m stations removed a substantial part of them, but not all. Daily scores of the first PC, $\gamma_{s,1}$, show extreme values during the days when a storm occurred. Therefore, they contain the residual signature of the storm which should be added to the index.

Let $p_{0.95,s}$ denote the 95^{th} percentile of the daily scores $\gamma_{s,1}$ for station s. We define

$$\gamma^*_{s,1} = \begin{cases} M_{\gamma_{s,1}}, & \text{if } |\gamma_{s,1}| > p_{0.95,s} \text{ for all } s, \\ \gamma_{s,1}, & \text{otherwise,} \end{cases} \tag{4.3}$$

where $M_{\gamma_{s,1}}$ is the median score of station s. This means that to extract the residual storm effect from the daily scores we find the largest 5% of the scores $\gamma_{s,1}$ for each station s individually. If the extreme value is captured by all stations we replace it by the median score, $M_{\gamma_{s,1}}$, of the corresponding station. The scores defined in (5.4) are used to compute daily periodic component P_s defined in (5.3).

Therefore, the residual storm contribution is

$$R_s(t) = D^c_{s,P}(t) - P_s(t), \quad t = 1, \cdots, N \tag{4.4}$$

It allows us to construct a so called pre-index, which consists of a storm signature extracted from three MRA levels of station s magnetogram. It is defined as follows

$$I_s(t) = \bar{D}_P(t) + R_s(t), \quad t = 1, \cdots, N, \tag{4.5}$$

and is the contribution of station s to the storm signature. An index is constructed by averaging the $I_s(t)$ from judiciously selected stations.

We now provide a brief summary of the procedure introduced in this section.

1. Perform the MRA on the raw megnetometer records, and extract details at levels $j = 8, 9, 10$ using MODWT and LA(8) filter. From here, work with the sum of these three levels of details.

2. Find the average of all stations used in the study, and remove it from the record at each station.

3. Convert the data from previous step into a functional one-day object. Compute the first FPC.

4. Replace the outlier scores aligned through all stations with median scores. Use those scores to estimate the daily component.

5. Compute the storm activity pre-index.

4.4 Comparison of Indices

The objective of this section is to compare the improved pre-index introduced in Section 4.3 to the known WISA pre-index, i.e. the part of the storm activity extracted from the three MRA levels.

First, we describe the datasets we use in this study. Then, we introduce a quantitative procedure we apply to compare the new pre-index to the WISA pre-index.

We use the H-component of the magnetometer records, the same as in WISA and Dst. Table 5.1 contains the list of the stations used in this study. To verify our results we use two four-station combinations and one six-station combination (see Table 4.2). The stations in each combination are roughly equispaced in the equatorial zone. The first set of four stations, HER, KAK, HON, SJG, was used because it is the standard Dst set, even though the stations HER, PHU , HON, SJG, are more equispaced. The index and the cannonical correlations for the two sets do not differ significantly. It is in fact an advantage of our method that it is fully automatic, and records from any set of stations can be used as inputs, and the resulting indices compared.

In order to produce a comprehensive study we applied the new procedure to 3, 5, and 6 month periods during 2001.

Figure 4.7 provides a visual comparison of the improved pre-index to the data, $D_{s,P}$, it was constructed from. During the disturbed periods of time (top panel) improved pre-index captures the storm signature (solid line), and most of the daily periodic component is eliminated from it during the quiet periods (bottom panel of Figure 4.7). What is left is in the range of ± 10 nT, slightly above the measurement error. During quiet days the original data are semi–periodic, so any index is semi-periodic, but with a smaller amplitude and often out of phase with the data.

A visual comparison is not enough to conclude that our proposed method removes daily periodic activity better than WISA or Dst. Therefore, we propose to use the functional canonical correlations to evaluate level of improvement quantitatively.

4.4.1 Functional canonical correlations

In this section we introduce the main idea of the functional canonical correlation analysis (FCCA).

Classical canonical correlation analysis (CCA) computes linear transformations of two variables such that the correlation between the transformed variables is maximized. Let (x_i, y_i), $i = 1, \cdots, n$, be pairs of observed vectors. The goal is to find vectors a_1 and b_1 such that the correlation between linear combinations $a_1' x_i$ and $b_1' y_i$ is the highest. A detailed discussion on CCA can be found in [43], Chapter 12.

Functional CCA provides a similar tool for investigating the relationship of the variability of functions. It helps to identify the modes of variability in the two sets of curves that are associated with each other most strongly.

Suppose, we observe N pairs of curves $(X_i(t), Y_i(t))$. Let (ξ, η) denote *canonical variate weight functions* defined such that the correlation between *canonical variates*

$\int \xi X_i$ and $\int \eta Y_i$ is the highest. As first observed by [44] and discussed in Chapter 11 of [4], to find a meaningful correlation an appropriate smoothing is essential. We are interested in comparing the canonical correlations for different methods and different datasets, therefore, we used the same smoothing parameter for all the data. The choice of the smoothing parameter does not change the overall conclusion.

In our application, the X_i represent the daily pre-index curves at one station (e.g HON) and Y_i the pre-index curves at another station (e.g HER). These curves can be computed using our method or WISA. If the canonical correlation for the curves computed using the new method is higher than for the curves obtained from the WISA procedure, it means that the new method isolates more features that are common for the X_i and the Y_i, e.g. for HON and HER.

The next section describes such comparison.

4.4.2 Quantitative comparison of different methodologies

In this section we present the results of the comparison of pre-indexes produced using the new improved method and WISA. We do not perform any direct comparison to the Dst. However, since WISA is statistically indistinguishable from Dst (see [19]), the conclusions we make about the WISA apply to the Dst index as well.

A method produces a "cleaner" pre-index (and index eventually) if local features are removed in a more efficient way. Therefore, the association of the pre-indexes of different stations should be strong. Using FCCA we show that our new technique isolates the effect of a storm better than the WISA procedure.

Denote by $\underline{D}_{s,8,W}$, $\underline{D}_{s,9,W,}$, $\underline{D}_{s,10,W}$ the details with the constant periodic component removed, as described in paragraph [42] of [19]. Thus,

$$R_{s,W} = \underline{D}_{s,8,W} + \underline{D}_{s,9,W,} + \underline{D}_{s,10,W} \tag{4.6}$$

is the WISA based pre-index.

We treat the pre-index I_s defined in (4.5) and WISA pre-index defined in (4.6) as functional data in UT. Let

$$\rho(s_1, s_2) = \text{ccorr}(I_{s_1}, I_{s_2}), \tag{4.7}$$

$$\rho_W(s_1, s_2) = \text{ccorr}(R_{s_1,W}, R_{s_2,W}) \tag{4.8}$$

be the canonical correlations between pre-indexes at stations s_1 and s_2 computed using improved methodology and the WISA approach.

Since the index of the storm activity is designed to capture global storm signature, the correlation between pre-indexes at different stations should be higher for the method that removes daily periodic component in a more efficient way. Figures 4.8 – 4.12 contain the canonical correlations for different combinations of stations during various periods of time. One can see that correlations for our proposed method $\rho(s_1, s_2)$ (star) are systematically higher than the correlations for WISA pre-index $\rho_W(s_1, s_2)$ (circle).

In order to check the effect of the averaging over all stations we constructed an alternative pre-index, I_s^*, where no averaging over all stations was performed, i.e. skipped step 2. The canonical correlations for this method are labeled with crosses in Figures 4.8 – 4.12. In most cases the resulting correlations are lower than the ones for WISA based pre-index. Therefore, we conclude that averaging over all stations effectively extracts storm signature.

4.5 Conclusions

We propose an improved procedure for removing the daily periodic component which uses statistical filtering techniques and functional principal component analysis

procedures. As an initial step we use multiresolution analysis to isolate the daily periodic component. To extract the storm signature we use the data from multiple stations. Principal component approach is applied to remove the non-constant daily variation. Our procedure produces an index which is cleaner than the WISA and the Dst both of which contain significant residual daily variation.

Functional canonical correlations were used to compute a quantitative measure of the level of improvement. We showed that there is a significant improvement from existing WISA and Dst, since WISA index is statistically indistinguishable from Dst. We conclude that our proposed methodology produces an index that isolates the global storm activity in a more efficient and cleaner way.

The method of deconvoluting the daily variation in the presence of a storm offers a potential tool to study its temporal and spacial behavior. Follow-on research will refine the technique of the analysis of the dynamic daily variation to study the relationship between the Sq, storm dynamo, and partial ring currents. Such an analysis is beyond the intended scope of this contribution, which emphasizes isolating global rather than local features.

Table 4.1: Geomagnetic observatories used in this study.

s	Name	Colatitude	Longitude
1	Hermanus (HER)	124.43	19.23
2	Antananarivo (TAN)	108.92	47.55
3	Phuthuy (PHU)	68.97	105.95
4	Kakioka (KAK)	53.77	140.18
5	Honolulu (HON)	68.68	202.00
6	San Juan (SJG)	71.89	293.85
7	Mbour (MBO)	75.62	343.03

Table 4.2: Combinations of four and six stations used to test the new method.

Set	Station
Four-1	HER, KAK, HON, SJG
Four-2	TAN, KAK, HON, SJG
Six-1	TAN, PHU, KAK, HON, SJG, MBO

Table 4.3: Combinations of four Dst stations (first set) used to compare methodologies.

Combination #	Stations
1	HON & KAK
2	HON & SJG
3	HON & HER
4	KAK & SJG
5	KAK & HER
6	SJG & HER

Table 4.4: Combinations of the second set of four stations used to compare methodologies.

Combination #	Stations
1	HON & KAK
2	HON & SJG
3	HON & TAN
4	KAK & SJG
5	KAK & TAN
6	SJG & TAN

Table 4.5: Combinations of the third set of six stations used to compare methodologies.

Combination #	Stations
1	HON & KAK
2	HON & SJG
3	HON & MBO
4	HON & TAN
5	HON & PHU
6	KAK & SJG
7	KAK & MBO
8	KAK & TAN
9	KAK & PHU
10	SJG & MBO
11	SJG & TAN
12	SJG & PHU
13	MBO & TAN
14	MBO & PHU
15	TAN & PHU

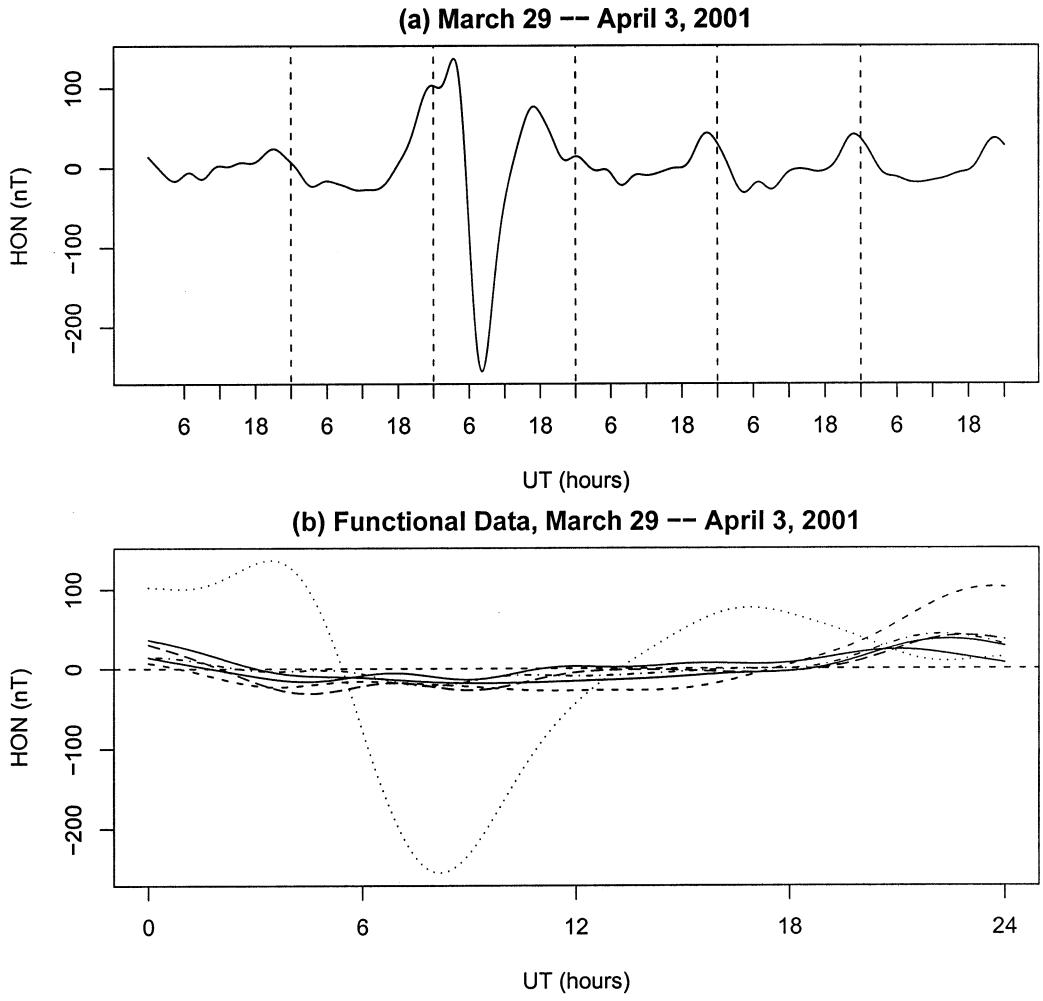


Fig. 4.1: (a) $D_{s,P}$ records, H-component during March 29 - April 3, 2001, HON, UT; (b) Functional data derived from the H-component during March 29 - April 3, 2001, HON, UT

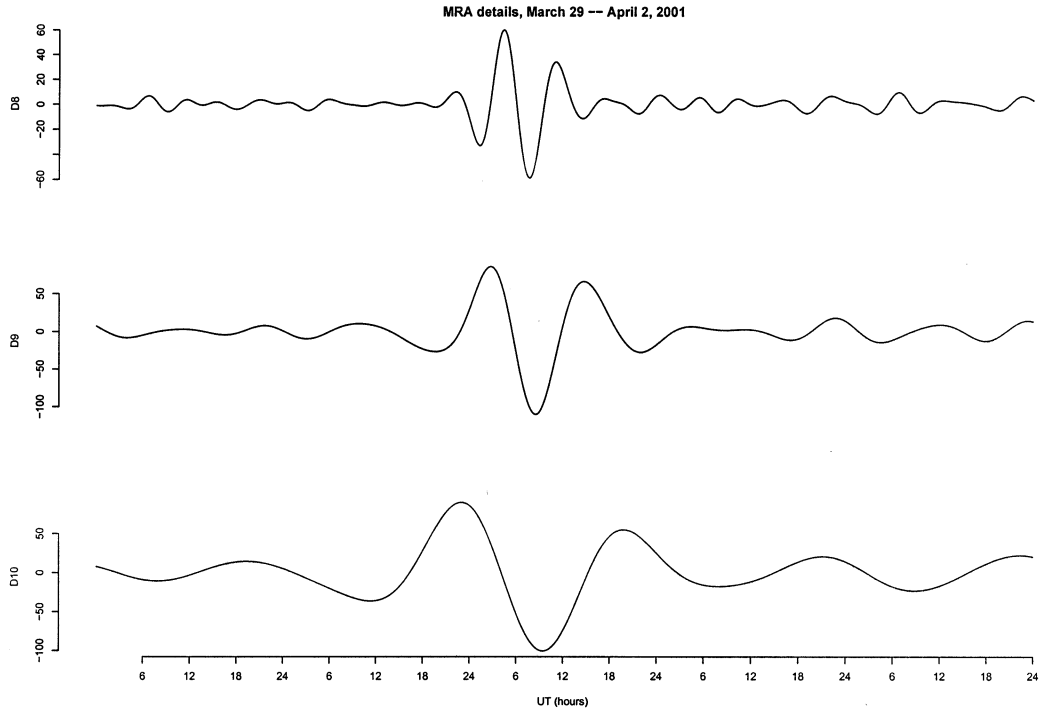


Fig. 4.2: Multi Resolution Analysis details D_8, D_9, D_{10}, March 29 - April 2, 2001, HON, UT

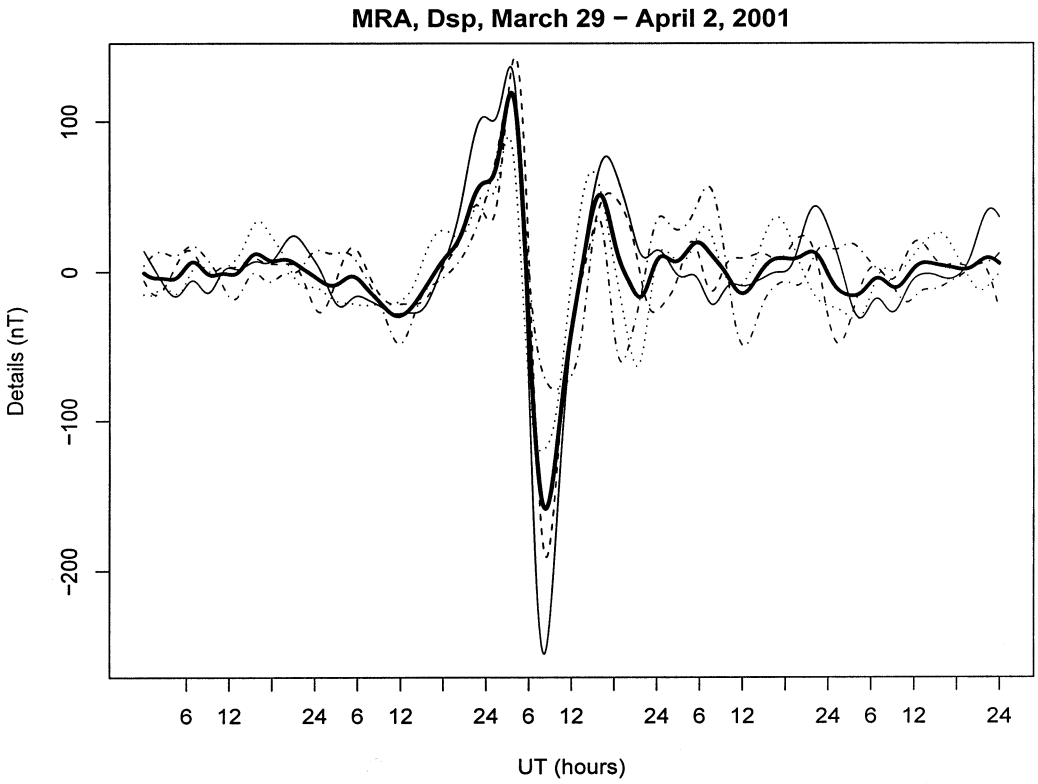


Fig. 4.3: $D_{s,P}$ components and their mean (thick line) of 4 Dst stations: HON, KAK, SJG, HER, during disturbed period of time: March 29 - April 2, UT

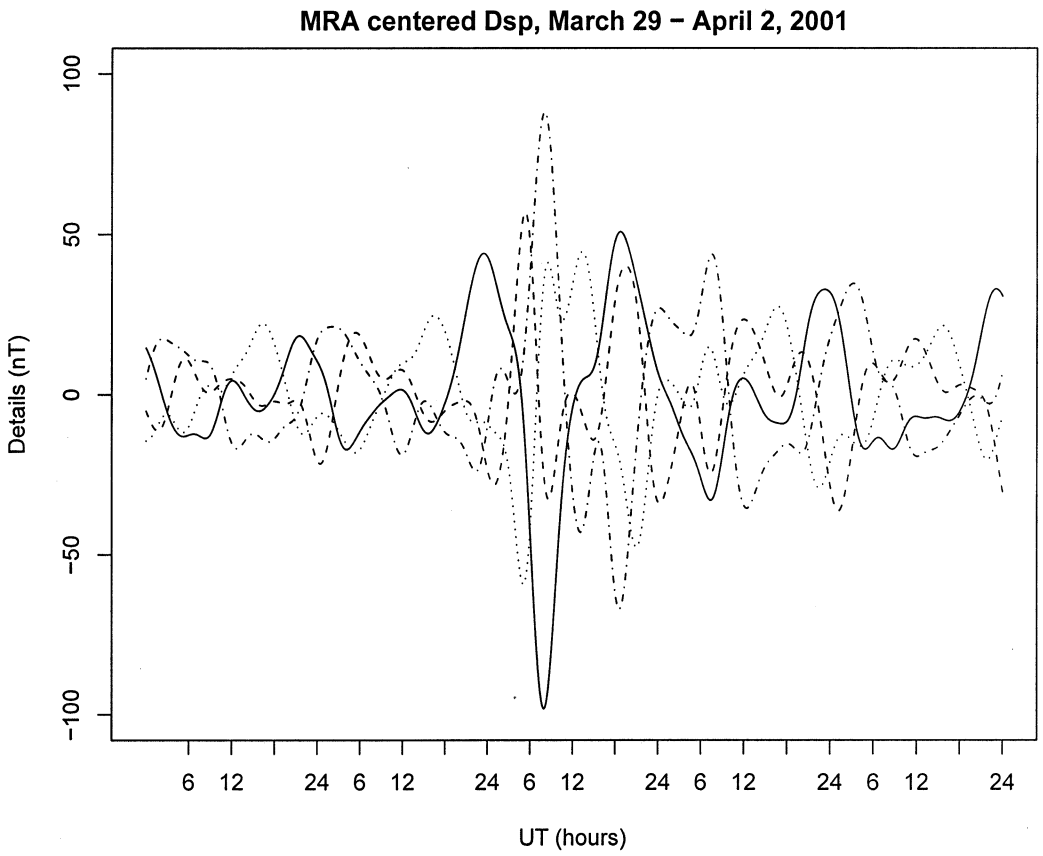


Fig. 4.4: Centered $D_{s,P}$ components of 4 Dst stations: HON, KAK, SJG, HER, during quiet time: March 29 - April 2, UT

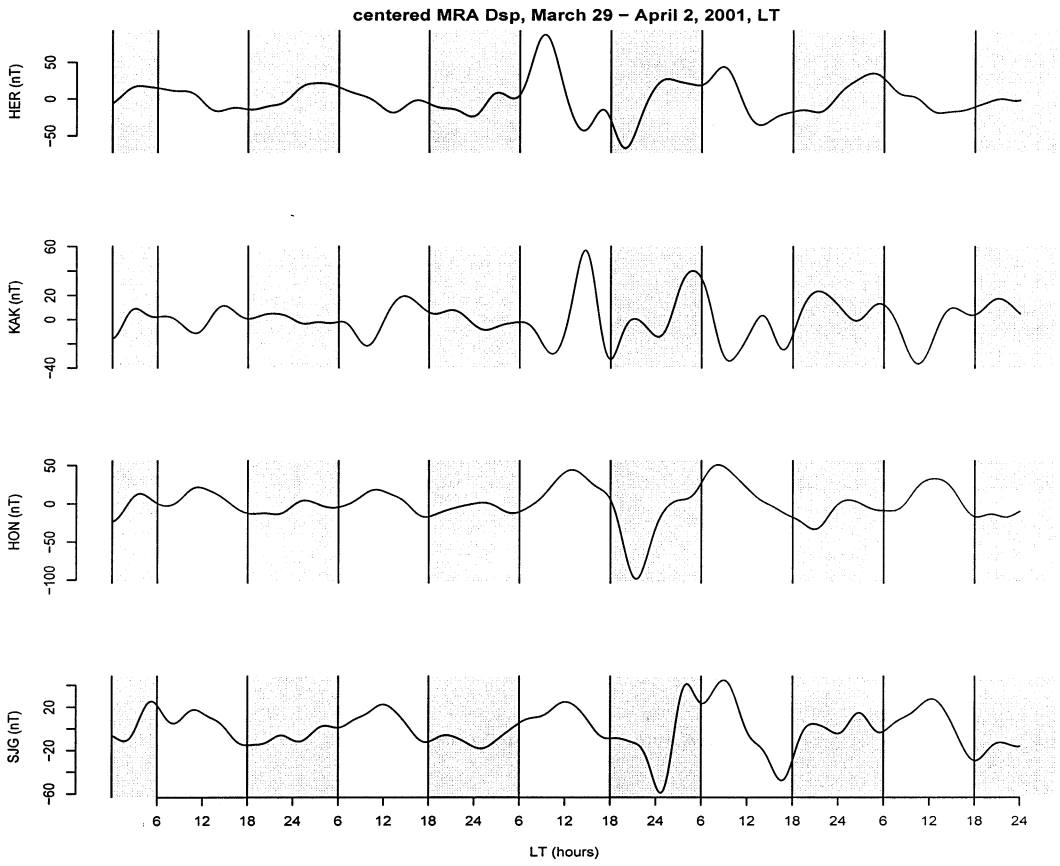


Fig. 4.5: $D^c_{s,P}$ components of 4 Dst stations: HON, KAK, SJG, HER, during disturbed period of time: March 29 - April 2, in LT. Grey areas correspond to night time

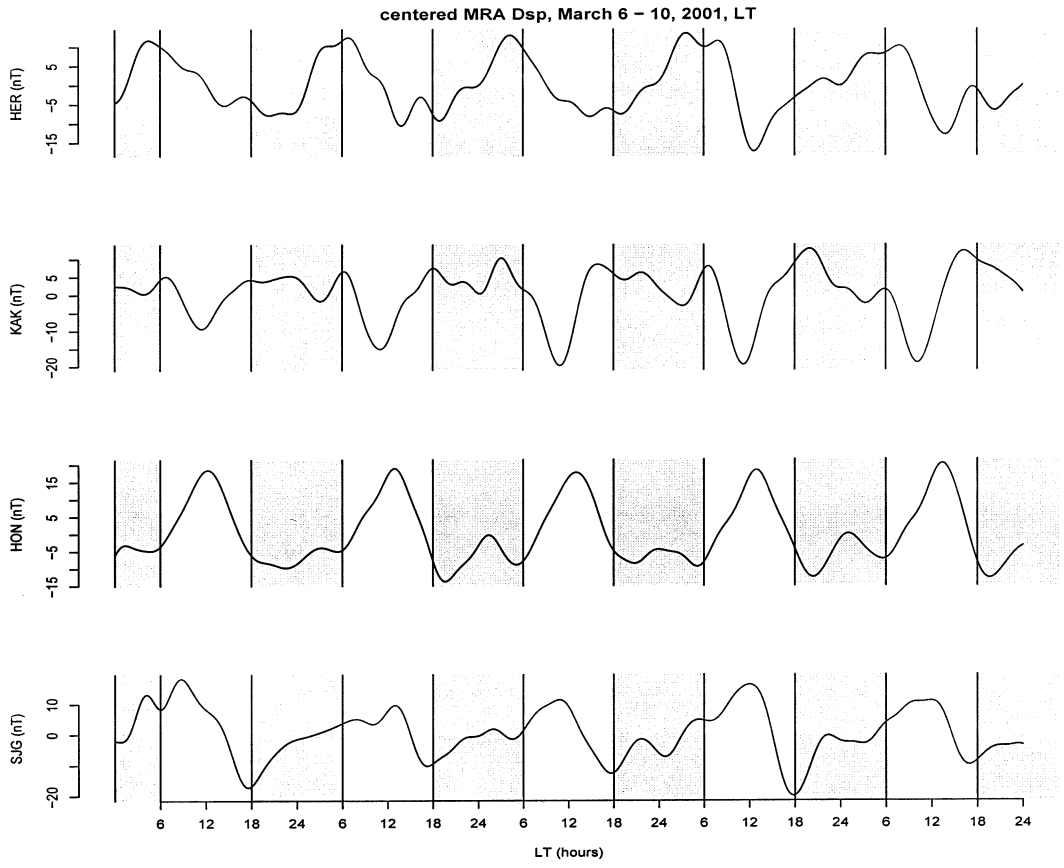


Fig. 4.6: $D^c_{s,P}$ components of 4 Dst stations: HON, KAK, SJG, HER, during quiet time: March 6 - 10, in LT. Grey areas correspond to night time

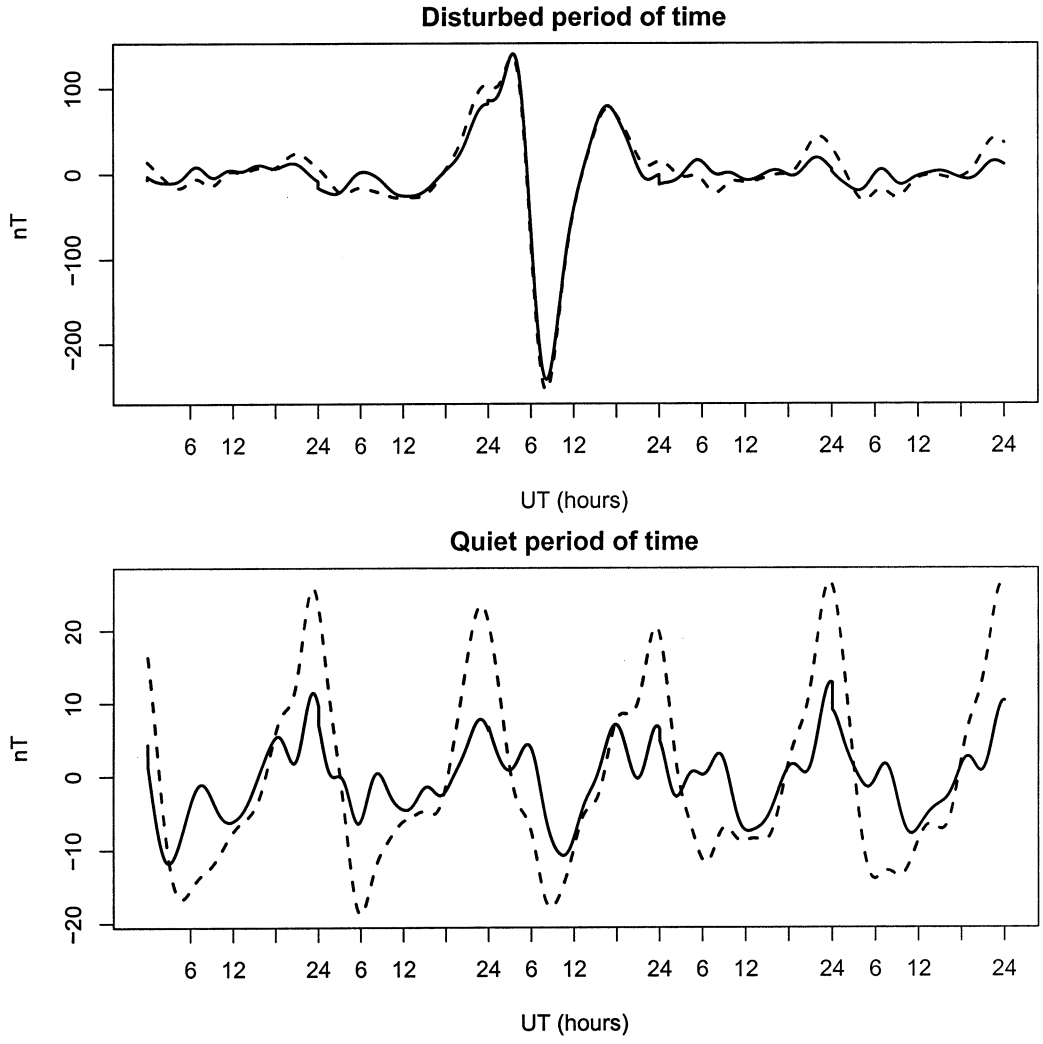


Fig. 4.7: Improved pre-index (solid line) and $D_{s,P}$ (dashed line) for HON station during disturbed (top panel) and quiet (bottom panel) periods

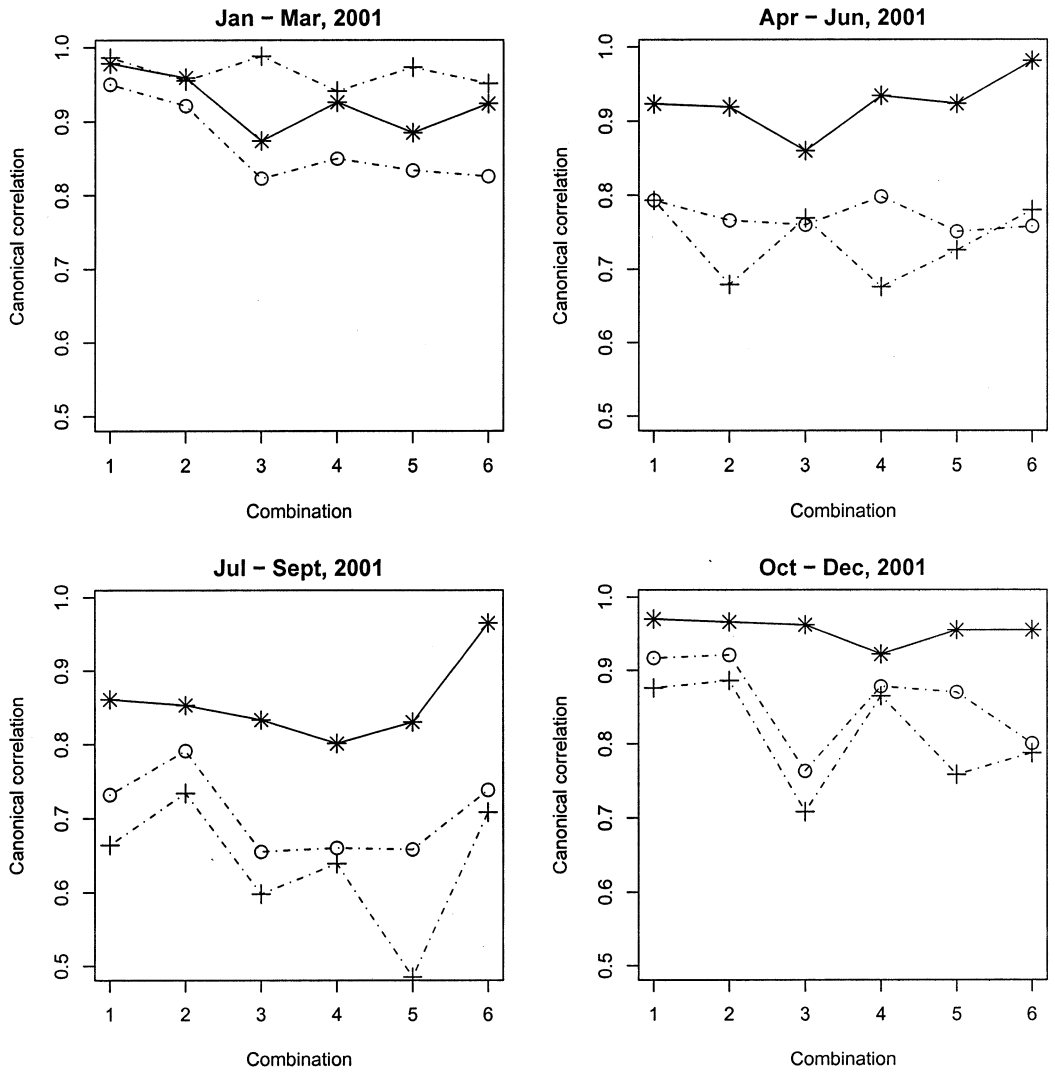


Fig. 4.8: Canonical correlations for the new method (star), new method without centering (cross) and WISA (circle), applied to all combinations of four Dst stations (see Table 4.3)

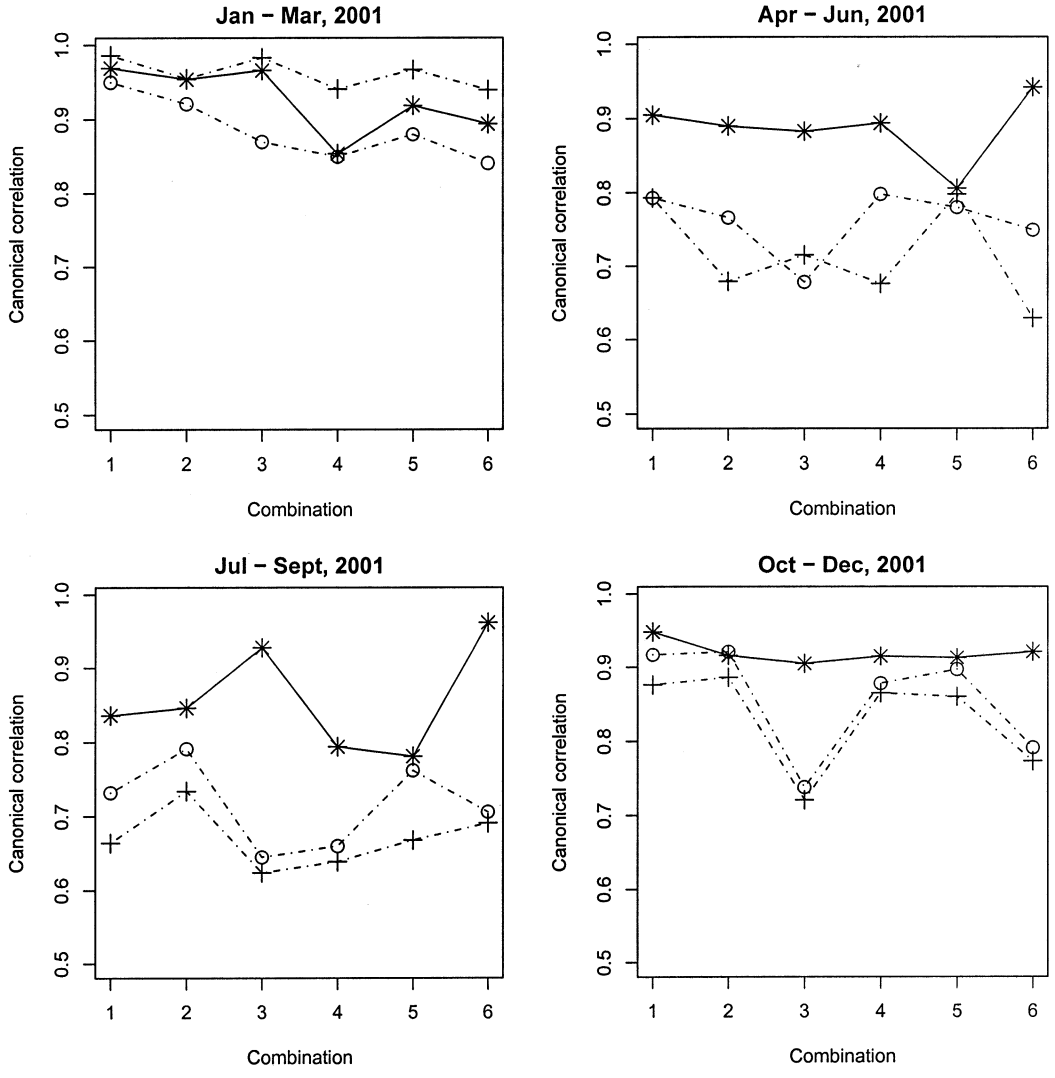


Fig. 4.9: Canonical correlations for the new method (star), new method without centering (cross) and WISA (circle), applied to all combinations of second set of four stations (see Table 4.4)

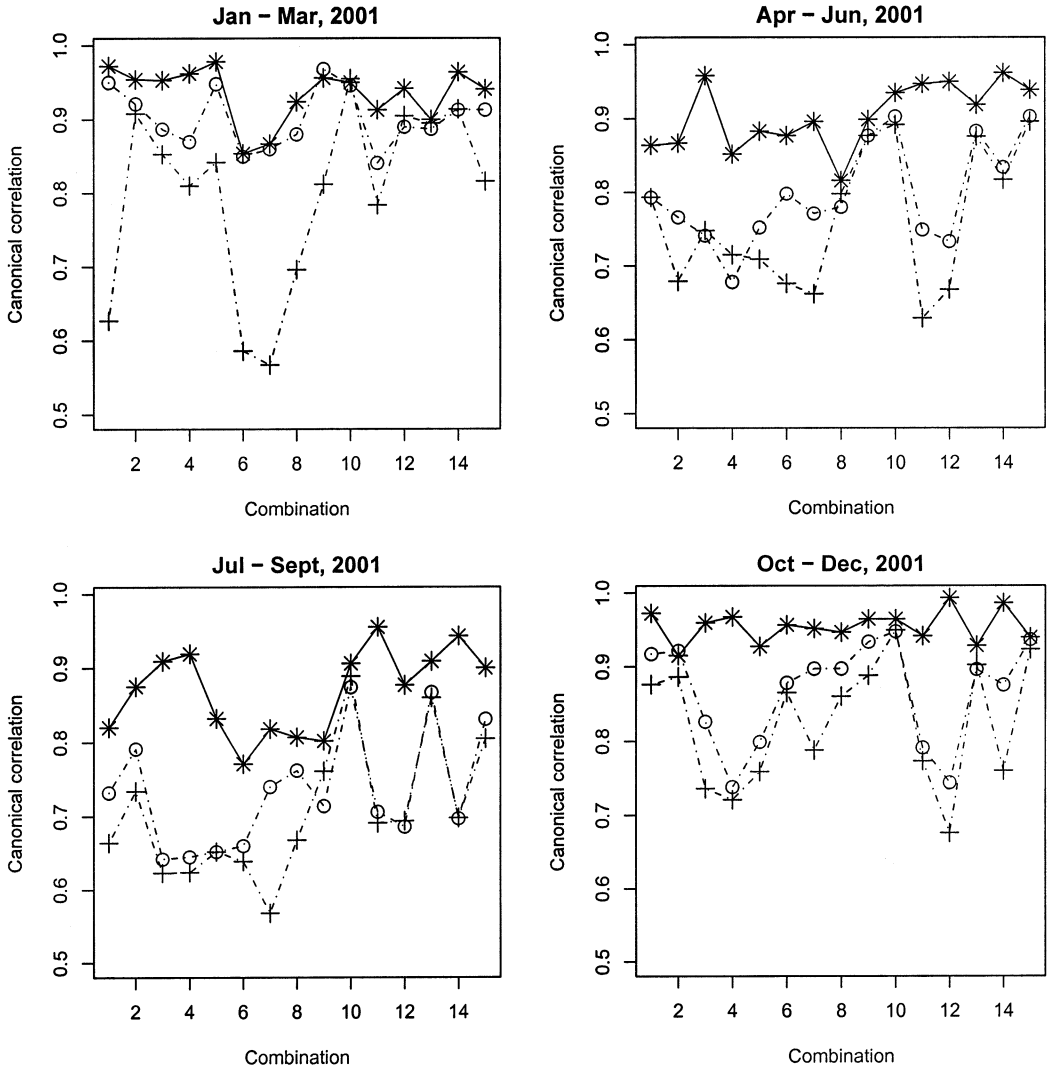


Fig. 4.10: Canonical correlations for the new method (star), new method without centering (cross) and WISA (circle), applied to all combinations of six stations (see Table 4.5)

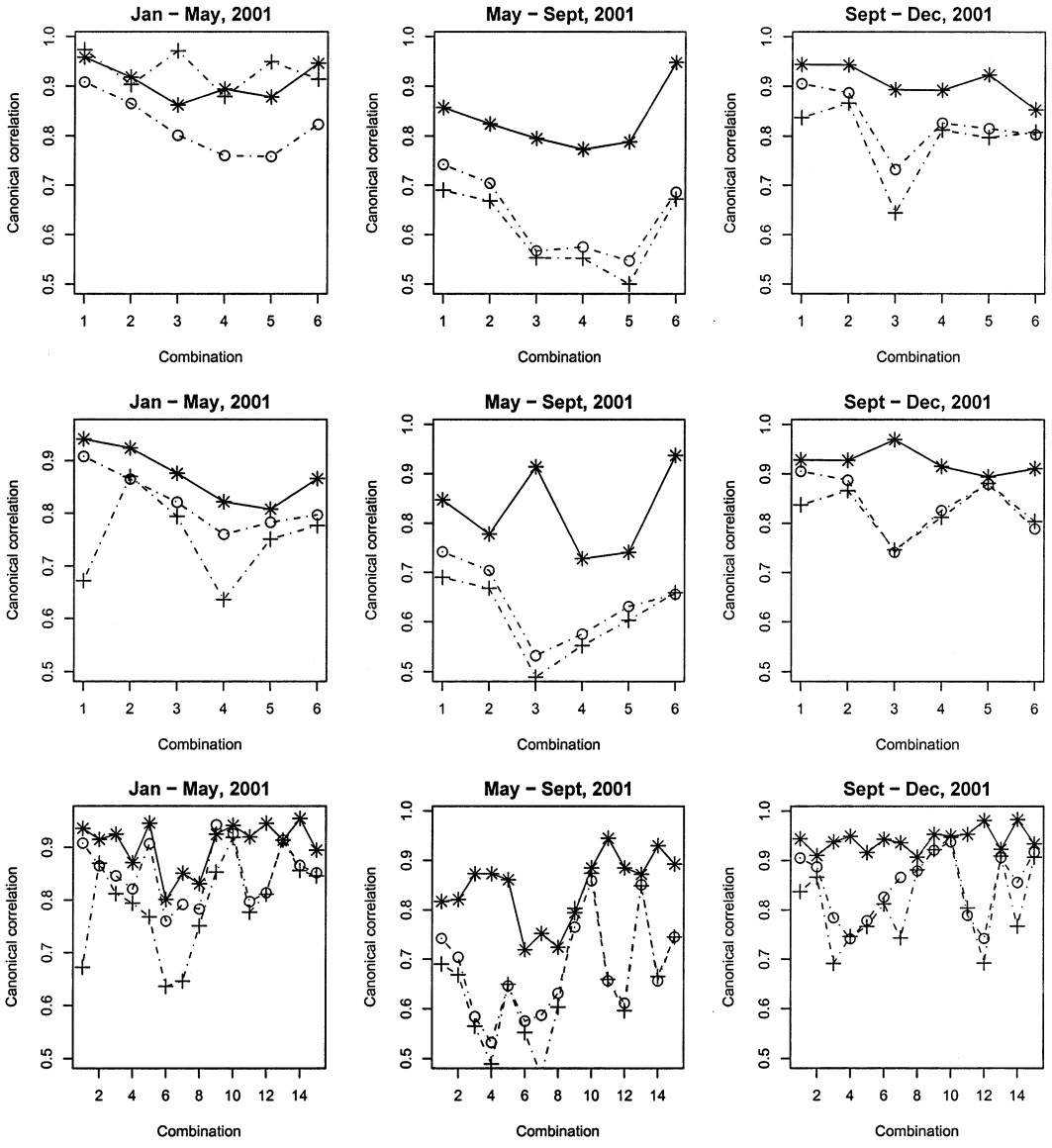


Fig. 4.11: Canonical correlations for the new method (star), new method without centering (cross) and WISA (circle), applied to all combinations of first set of stations (top panel, combinations are given in Table 4.3), second set of stations (middle panel, combinations are given in Table 4.4), third set of stations (bottom panel, combinations are given in Table 4.5)

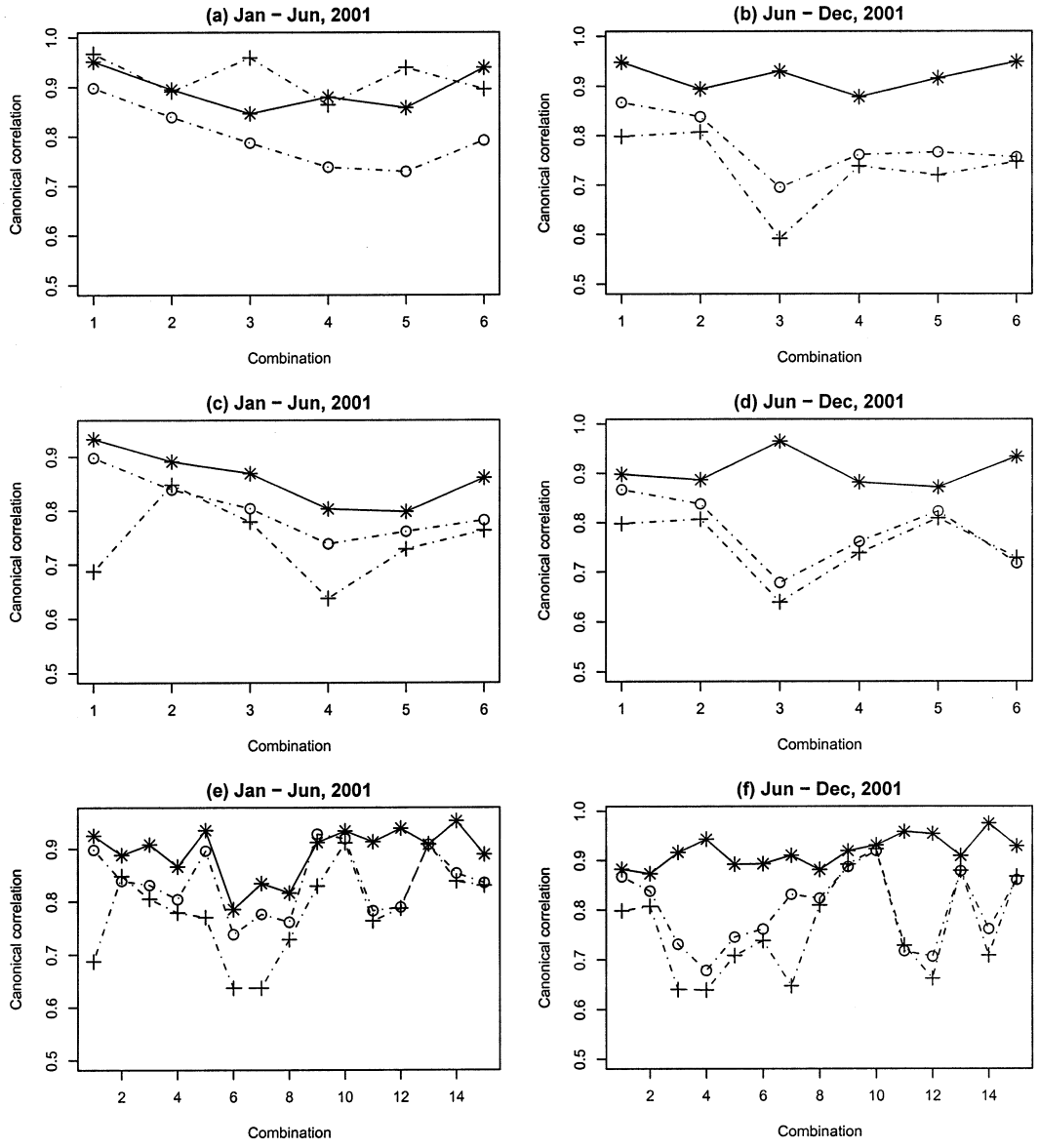


Fig. 4.12: Canonical correlations for the new method (star), new method without centering (cross) and WISA (circle), applied to all combinations of first set of stations (top panel, combinations are given in Table 4.3), second set of stations (middle panel, combinations are given in Table 4.4), third set of stations (bottom panel, combinations are given in Table 4.5)

CHAPTER 5
ESTIMATION OF SQ VARIATION BY MEANS OF MULTIRESOLUTION AND PRINCIPAL COMPONENT ANALYSES[1]

5.1 Introduction

We propose a new method of estimating the Sq component of low latitude magnetometer records. While our method builds on the method recently proposed by [37], it is different in that it uses multiple stations and wavelet filtering.

An important part of the algorithm used to compute the Dst ([45]) is the estimation the quiet daily variation, which is essentially defined as the average of a few quiet days in a month. The Sq component calculated this way is the same for all days in a month, and it has been recognized that such an assumption is not accurate. Even on quiet days, the daily variation changes very visibly from day to day, both in its amplitude and its shape. This is attributable to multiple dynamic drivers which include not only tidal ionospheric winds, but also the effect of the Chapman-Ferraro current, the Sq current, and the magnetotail current, [18], [37], and references therein. On storm days, the interactions of these drivers are even more complex.

We propose a technique of isolating the low latitude Sq variation which relies on wavelet and functional data analysis methods, and which is a refinement of an algorithmic step in the construction of an index of symmetric equatorial storm activity developed in [20]. While the focus of the procedure of [20] was on extracting symmetric global features mainly attributable to the (enhanced) ring current, the present paper aims at isolating LT features attributable to ionospheric winds driven by solar heating. We do not estimate LT features due to the storm time enhancements of

[1]Coauthored by I. Maslova, P. Kokoszka, J.J. Sojka, and L. Zhu.

the partial ring current. A good technique for isolating Sq features should produce curves with a high degree of appropriately measured similarity (in LT) for neighboring stations. The LT H-component looks very different at different stations, but during the same UT day these different shapes are formed by approximately the same solar drivers, so the "correlation" of LT curves approximating the Sq for neighboring stations should be high. On the other hand, changes symmetric in UT, should not be reflected in the Sq estimates. These are our guiding principles for the construction of Sq estimates.

The traditional method of computing the quiet daily variation consists in finding 5 most quiet UT days (as measured by an index like Kp), averaging the H-components over these five days and smoothing. This results in a daily curve which differs across stations, but is constant for every day in a month at a given station. This approach has recently been improved upon by [19] and [46] who use very different approaches and motivation, but produce estimates of the daily variation with very similar properties. [19] focus on equatorial stations, and define the daily component as the median of the sum of three levels in a wavelet multiresolution analysis (MRA). The appropriate levels of the MRA isolate the time scales characteristic of equatorial Sq, whereas the median produces a typical daily shape which is not affected by outliers (large disturbances). By using a moving window of flexible length (e.g. 30 days), an Sq that changes slowly from day to day can be calculated. [46] focus on auroral currents and use a more complex smoothing and outlier removal technique to determine a gradually changing quiet daily variation using a moving window of 30 days. By imposing certain conditions on the smoothness of the quiet daily curve, they can algorithmically identify quiet days. The time scales involved in their approach are 20 minutes and 3 hours, and, appropriately to the task (auroral activity), are much smaller than the time scales used by [19]. Both techniques are fully algorithmic, and

produce a slowly changing pattern (the curves on two consecutive days look almost identical). [47] develop an algorithm for predicting such patterns, which incorporates the long term secular trend, the solar cycle and other long term variations. Their data are 24 time series (one for each UT hour) collected over a period of 70 years, producing 840 observations (one per month) for each of the 24 series.

Our goal is not the long term modeling of the evolution of a "typical" Sq pattern. We aim at good estimates of dynamic Sq over relatively short intra-annual periods capable of reflecting its often spectacular day–to–day variability [48]. By providing free software which implements our procedure, available at *http://wami.usu.edu*, we would like to offer a new tool for the community. An example of the variable Sq obtained using our software is shown in Figure 5.1.

First, in Section 5.2 we describe measures of similarity of curves which we use to assess the quality of Sq estimates. We illustrate their applicability to synthetic Sq curves in Section 5.3. In Section 5.4, we describe the proposed new method in detail, and contrast it with the method of [37]. Section 5.5 focuses on the comparison of the two methods, while Section 5.6 concludes.

5.2 Measures of Time–Aligned Similarity of Curves

In this section we discuss measures of curve similarity that we use to compare methods of Sq estimation.

Correlations are often used as a measure of association. Let $Q^{(1)}$ and $Q^{(2)}$ be two samples of size N. The sample mean is $\bar{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i$, and sample variance is $s_Q^2 = \frac{1}{N-1}(Q-\bar{Q})^2$. The sample correlation between $Q^{(1)}$ and $Q^{(2)}$ is

$$r_{Q^{(1)},Q^{(2)}} = \frac{\sum_{i=1}^{N}(Q_i^{(1)} - \bar{Q}^{(1)})(Q_i^{(2)} - \bar{Q}^{(2)})}{(N-1)s_{Q^{(1)}}s_{Q^{(2)}}}.$$

Correlation is affected by extremely large values (outliers, see e.g. Chapters 8 and 9 of [49]). In the context of this paper, such extreme values can occur when the Sq estimate contains storm related features, which would result in a highly correlated data. We therefore propose a new measure of similarity of two samples of curves.

Let the first sample is denoted $Q_n^{(1)}(t),\ t \in [0,T],\ n = 1,2,\dots N_D$, the second $Q_n^{(2)}(t),\ t \in [0,T],\ n = 1,2,\dots N_D$. In the context considered in this paper, the interval $[0,T]$ represents an LT day, and N_D the number of days in each sample.

We define

$$D_n(Q_n^{(1)}, Q_n^{(2)}; a) = \frac{1}{T}\sum_{t=1}^{T} |Q_n^{(1)}(t) - aQ_n^{(2)}(t)|,\ \ n = 1,2,\dots N_D$$

and

$$\hat{D}(Q^{(1)}, Q^{(2)}) = \min_a \frac{1}{N_D}\sum_{n=1}^{N_D} D_n(Q_n^{(1)}, Q_n^{(2)}; a). \tag{5.1}$$

The measure $\hat{D}(Q^{(1)}, Q^{(2)})$ is small if the two samples have similar time aligned features. Unlike correlation, $\hat{D}$ is not shift invariant, i.e. if a constant is added to all curves in one of the samples, the value of $\hat{D}$ will change. This is a desirable property because base line fields should be removed form estimates of Sq variations. The two samples do not have to be identical for $\hat{D}$ to be zero. If for each n, $Q_n^{(2)}(t) = cQ_n^{(1)}(t)$, for some constant c, then $\hat{D}(Q^{(1)}, Q^{(2)}) = 0$.

In addition to the measure of the curve similarity introduced above we use the wavelet power spectrum to analyze the estimated Sq.

Let $Q_t, t = 0,1,\cdots,N$ be the one minute estimate of the Sq of length N.

The discrete wavelet power spectrum associated with a scale $\tau_j = 2^{j-1}$, where $j = 1, 2, \cdots, J$ is

$$P_{\tilde{W}}(\tau_j) = \frac{1}{N}\|\tilde{\boldsymbol{W}}_j\|^2,$$

where $\tilde{\boldsymbol{W}}_j$ is a vector the wavelet coefficients obtained filtering Q_t (see Chapter 5 of [42] for more details). The empirical power spectrum is the variance of the wavelet coefficients. We use it to see which frequency of the estimated Sq is most pronounced. The values of $P_{\tilde{W}}(\tau_j)$ should be higher for the levels j that capture daily variations.

We test measures of curve similarity using synthetic Sq estimates introduced in the following Section.

5.3 Application to Synthetic Sq Curves

In this section we construct synthetic examples of "good" and "bad" Sq estimates, and apply to them the curve similarity measures described in Section 5.2. We simulate several pairs of synthetic Sq estimate curves (see Figures 5.2 and 5.3). Assume that these pairs of curves are the estimates found using data from neighboring stations aligned in local time (LT). The "bad" Sq estimate includes some extreme values or even patterns that are present at all stations and aligned in UT. It means that the storm features were not completely removed. The "good" Sq estimate consists of the daily pattern that stays roughly the same each day. The amplitude of such an estimate varies slightly from day to day. Therefore, it is an estimate of a nonconstant SQ. Finally, there are no global features present.

Define the synthetic daily pattern as

$$sQ(t) = 0.6\sin\left(\frac{\pi t}{1440}\right) + 0.2\sin\left(\frac{2\pi t}{1440}\right), \qquad t = 1, \cdots 1440,$$

Next, we generate a "good" Sq estimate $Q^{(1)}_{s,n}(t)$, for which the main features are adjusted in local time. Let

$$Q^{(1)}_{s,n}(t) = \{sQ(t) + R_s(t)\}\, U_n,$$

where U_n is a uniformly distributed on $[0.5, 1.5]$ daily noise that is the same for all stations s. Here,

$$R_s(t) = 0.5 V_{s,n} \sin\left(\frac{2\pi t}{1440}\right),$$

where $V_{s,n}$ is uniformly distributed on $[0, 1]$ noise. This is an additive daily noise function that is different for all stations s.

Thus, $Q^{(1)}_{s,n}(t)$ simulates a "good" non-constant Sq estimate that captures daily pattern aligned in local time. Day-to-day variability is introduced by the noise variables U_n and $R_s(t)$ (see Figure 5.2).

Next, we present an example of a "bad" non-constant Sq estimate. Define

$$G(t) = 2\sin\left(\frac{2\pi t}{1440}\right) I_{[0;720]}.$$

Let

$$G_s(t) = G(t + \Delta_s),$$

where $\Delta_1 = 0$, $\Delta_2 = 360$, $\Delta_3 = 520$, $\Delta_4 = 720$. Here, Δ represents the shift of the pattern that occurs when some UT features are not removed properly.

Set

$$Q^{(2)}_{s,n} = \{sQ(t) + R_s(t) + G_s(t)X_n\}\, U_n,$$

where X_n is a Binomial random variable with $p = 0.3$.

Figure 5.3 presents an example of a "bad" Sq estimate where the extreme spikes

are not aligned in time. The same happens when the global storm features are not removed from the quiet day component estimate.

Next, we simulate both "good",$Q^{(1)}_{s,n}$, and "bad", $Q^{(2)}_{s,n}$, Sq estimate curves for four stations during 15 days, i.e. $s = 1, \cdots, 4$, $n = 1, \cdots, 15$. These curves are used to evaluate the curve similarity measures introduced in previous section. The associations between simulated data sets for different stations are roughly linear, therefore the correlation analysis is appropriate. Panel (a) of Figure 5.4 provides the correlations for two simulated Sq estimates. Since we generate data for four stations we get six different combinations, hence, six correlation values for each Sq method. We conclude that correlations separate these two synthetic Sq estimates. Correlations capture the shifts in time Δ, but are not sensitive towards changes in the amplitude.

However, in case of real data simple correlation does not perform as good as for simulated data. First, the association of Sq extracted from real magnetometer data from different stations is not linear. Second, if the method of Sq estimation fails to remove most of the storm features the correlation values become very high. In that case the high correlation is due to the extremely large values in Sq estimate which are attributable to the remaining storm signature. Therefore, a clearly "bad" Sq estimate gives high correlation values. Panel (b) of Figure 5.4 shows that minimized average distance clearly distinguishes between the "good" and "bad" Sq estimates.

5.4 Estimation of a Non-Constant Solar Quiet Daily Variation

A method of natural orthogonal components, which we call here the principal components, to identify the solar quiet variation is introduced by [50]. Like for [37], this observation is the starting point of our method. We, however, introduce a number of important refinements.

First, a wavelet-based representation of the data is introduced. Let, $X_s =$

$\{X_{s,t}, t = 1, \cdots, N\}$ be the magnetogram recorded at station $s = 1, \cdots, m$, where N is the length of the record in minutes, e.g. two months. We can write it as

$$X_s(t) = \sum_{j=1}^{J} D_{s,j}(t) + S_{s,j}(t),$$

where $D_{s,j} = \{D_{s,j}(1), \cdots, D_{s,j}(N)\}$ are the details, and $S_{s,j} = \{S_{s,j}(1), \cdots, S_{s,j}(N)\}$ is the smooth. Here, $j = 1, \cdots, J$ is the multiresolution analysis (MRA) level. The details capture the part of the records that corresponds to the frequencies in the range from 2^{-j-1} to 2^{-j} cycles per minute. For further details see Chapter 5 of [42].

The Sq component is most clearly pronounced in the MRA details $D_{s,j}$ for levels $j = 8, 9, 10$. These levels capture different parts of the Sq spectrum; level $j = 8$ captures approximately the 6 hour periodic component, $j = 9 - 12$ hour component, and $j = 10 - 24$ hour component. However, these details are enhanced during a storm. Comparing the records from several stations we can clearly see that these disturbances are aligned in UT. Therefore, they are not the part of Sq and should be removed. To remove storm associated features we use the storm index introduced in [19] and improved by [20]. Four Dst stations are used to construct the storm activity index (see Table 5.1). However, one can use any roughly equispaced equatorial stations. The stations used for storm index estimation must not include the stations where Sq is to be estimated.

Let, $I(t), t = 1, \cdots, N$ be the storm index of [20] reflecting the strength of the ring current. We remove $I(t)$ from the H-component of all the stations used in the study.

After removing the storm index from the data, we perform the multiresolution analysis (MRA), using maximum overlap discrete wavelet transform MODWT and LA(8) filter. We can write it as

$$X_s(t) - I(t) = \sum_{j=1}^{J} D_{s,j}(t) + S_{s,j}(t), \quad t = 1, \cdots, N.$$

Let

$$D_{s,Q}(t) = D_{s,8}(t) + D_{s,9}(t) + D_{s,10}(t), \quad t = 1, \cdots, N$$

be the part of the signal that includes practically all frequencies of the daily component spectrum. The subscript "Q" stands for "quiet"daily component.

We therefore postulate that

$$D_{s,Q}(t) = Q_s(t) + R_s(t), \quad t = 1, \cdots, N, \tag{5.2}$$

where Q_s is identified with the solar quiet daily periodic component and R_s is the residual effect of a storm. Next, we apply principal component analysis techniques to estimate the daily variation Q_s. We convert $D_{s,Q}$ into functional object, i.e. daily functions that start at UT midnight. Using principal component analysis we can write (t' is the time in minutes within one day)

$$D_{s,Q}(t') = \mu_s(t') + \sum_{j=1}^{\infty} \gamma_{s,j} u_{s,j}(t'), \quad t' = 1, \cdots, 1440,$$

where $\mu_s(t')$ is the daily mean, $\gamma_{s,j}$ is a score vector for j^{th} PC, and $u_{s,j}$ is the j^{th} PC for station s.

Denote the number of days by $N_D = N/1440$. We assume that the periodic component for day $i = 1, \cdots, N_D$ is

$$Q_{s,i}(t') = \mu_s(t') + \gamma^*_{s,1,i} u_{s,1}(t'), \quad t' = 1, \cdots, 1440, \tag{5.3}$$

where $\mu_s(t')$ is the daily mean, $\gamma^*_{s,1,i}$ is a filtered score for the i^{th} day described below.

The function $u_{s,1}(t')$ is the first PC for station s. In (5.3) $\mu_s(t')$ and $u_{s,1}(t')$ are deterministic functions defined over the 24-hour interval, and $\gamma^*_{s,1}$ are random weights that change from day to day. Hence, the extracted Sq, $Q_s(t)$, is non-constant. Note that $Q_{s,i}(t')$ where $t' = 1, \cdots, 1440$ and $i = 1, \cdots, N_D$ is the same daily periodic component as $Q_s(t)$ where $t = 1, \cdots, N$ split into daily functions.

Even after removing the storm signature from the magnetometer records the three selected MRA levels may still contain residual storm features. Daily scores of the first PC, $\gamma_{s,1}$, show extreme values during the days when a storm occurred. Therefore, they contain the residual signature of the storm which should be removed.

Let $M_{s,1} = \text{median}(\gamma_{s,1,i}, i \leq N_D)$ be the median first principal component score for station s . Further, let $p_{0.90,s}$ denote the 90^{th} percentile of the absolute value of the daily median adjusted scores for station s, i.e. $|\gamma_{s,1} - M_{s,1}|$. We define

$$\gamma^*_{s,1} = \begin{cases} M_{s,1}, & \text{if } |\gamma_{s,1} - M_{s,1}| > p_{0.90,s} \text{ for all } s, \\ \gamma_{s,1}, & \text{otherwise,} \end{cases} \tag{5.4}$$

where $M_{\gamma_{s,1}}$ is the median score of station s. This means that to eliminate the residual storm effect from the daily scores we find the largest 10% of the scores $|\gamma_{s,1} - M_{s,1}|$ for each station s individually. If the extreme value is captured by all stations we replace it by the median score, $M_{s,1}$, of the corresponding station. The scores $\gamma^*_{s,1}$ defined in (5.4) are used to compute the daily periodic component P_s defined in (5.3). In order to get better Sq estimate it is necessary to use records from at least two stations. Otherwise there is no way to eliminate the residual global features, as described above.

Decomposition (5.3) is akin to the ideas of [18] and [37], who argued that the first principal component follows the pattern of the daily Sq-variation. However, while

these authors work with the raw magnetometer records, we first remove the global storm signature to eliminate the storm effect and then apply a wavelet filter to the data and use just the levels that contain the periodic component. So, in our paper, to estimate daily periodic component Q_s, we eliminate the global storm features that are not the part of Sq. We compute the first PC of $D_{s,Q}$ rather than the first PC of the raw magnetometer data with some seasonal adjustments which do not remove the storm activity from the Sq. Our method combines data from multiple low latitude stations.

5.5 Comparison of the Sq Estimates

The goal of this section is to provide a detailed comparison of the Sq estimate introduced in Section 5.4 and the method proposed by [37]. We refer to the approach proposed here as the *new method* and the approach of [37] as the *alternative method.*

First, we introduce the data used for this comparison. We use the H-component of the ground-based magnetometer records. Table 5.1 provides the list of the geomagnetic stations used in our study. As mentioned in Section 5.4, first, we remove the storm index from the raw data. The storm index is computed from four Dst stations: Hermanus, Kakioka, Honolulu, San Juan.

We compare the estimates of the daily variation for the following pairs of stations: (1) Alibag (ABG) and Phuthuy (PHU), (2) Tucson (TUC) and Fredericksburg (FRD). Note that the stations in each pair are relatively close to each other, so that the data used for Sq extraction is generated by the same ionospheric configuration.

We present the results for two periods of time: a one month period, February 2001, and two month – March - April, 2001. The main difficulty of the Sq estimation occurs during geophysically disturbed time, hence the choice of the time intervals. February 2001 is a relatively quiet period of time, the second period, March – April,

contains several extremely strong storms, one of them took place on March 31, 2001 (see Figure 5.6).

Figure 5.7 provides a visual comparison the two Sq estimates to the raw magnetometer records. One can see that the new Sq (dashed line) follows the quiet daily pattern (solid line) for both quiet and disturbed times. However, the alternative Sq (dotted line) is very affected by the extreme values of the storm signature. During the storm it drops up to -400nT, which definitely contradicts to the Sq definition. The new procedure eliminates the global storm effects, therefore, the Sq remains stable during very strong storms with only a mild enhancement during the sudden commencement phase (see bottom panel of Figure 5.7). The poor performance of the alternative method during the quiet period shown in the top panel of Figure 5.7 is due to the presence of storms during the two month period used to construct these Sq estimates. If a shorter period without strong storms is used for the estimation then the alternative method gives results comparable to the new procedure. Long quiet periods are however rare.

As part of the analysis, we compare the wavelet-based power spectrums of the Sq obtained using our technique, the alternative methodology, and raw data. Figure 5.8 presents an example of the empirical wavelet power spectrum based on MODWT for $j = 1, \cdots 13$. Note that the daily variations are captured by levels $j = 8, 9, 10$. We can see that the raw data (circles) have a significant daily component present, as well as larger scale variations. A good Sq estimate should mostly capture the daily periodic component. The higher and lower frequency variations should be insignificant. One can see that our Sq estimate captures the daily variability, and the largest spectrum values are at $j = 8, 9, 10$, which is a desirable property (see the line with stars in Figure 5.8). The spectra of the alternative Sq estimate roughly follows the raw data spectra. The daily component is not significant (see the line with crosses in

Figure 5.8). The highest values are observed for levels $j > 10$, which are associated with variations on scales larger than the daily scale.

Finally, we find the measure of the curve similarity $\hat{D}$ for two pairs of the stations used in our study. Table 5.2 provides the quantitative results. The minimized average distance between the Sq estimates at neighboring stations is smaller for the technique introduced in this paper. That means that our proposed procedure extracts the local time aligned features better.

5.6 Conclusions

We introduce an automated procedure of extracting the Sq signature from the H-component of low latitude magnetometer records. Wavelet and functional analysis techniques are applied. The methodology proposed here uses the index of storm activity to remove the main storm features as an initial step. In order to extract the daily variation we use the data from multiple stations. Multiresolution analysis is used to better isolate the part of the signal that captures the Sq component. We use the functional principal component approach to estimate the nonconstant daily variation. Our methodology gives the Sq estimate that extracts the local time features in a more accurate way and is more stable than other methods.

Table 5.1: Geomagnetic observatories used in this study. Stations used to estimate the ring current activity are labeled with *.

s	Name	Colatitude	Longitude
1	Hermanus* (HER)	124.43	19.23
2	Alibag (ABG)	71.38	72.87
3	Phuthuy (PHU)	68.97	105.95
4	Kakioka* (KAK)	53.77	140.18
5	Honolulu* (HON)	68.68	202.00
6	Tucson (TUC)	57.82	249.27
7	Fredericksburg (FRD)	51.80	282.63
8	San Juan* (SJG)	71.89	293.85

Table 5.2: Distance $\hat{D}$ (5.1) between the estimated Sq curves.

Method	February		March – April	
	ABG & PHU	TUC & FRD	ABG & PHU	TUC & FRD
New Sq	3.84	4.29	5.13	6.81
Alternative Sq	6.06	5.92	10.42	9.40

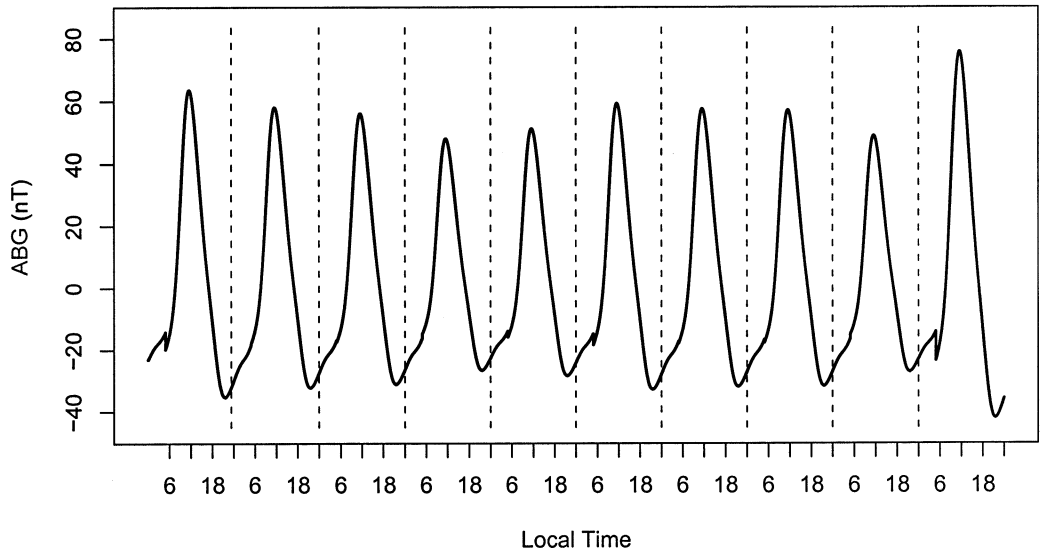


Fig. 5.1: Estimated Sq, Alibag (ABG) station during March 21 – March 30, 2001.

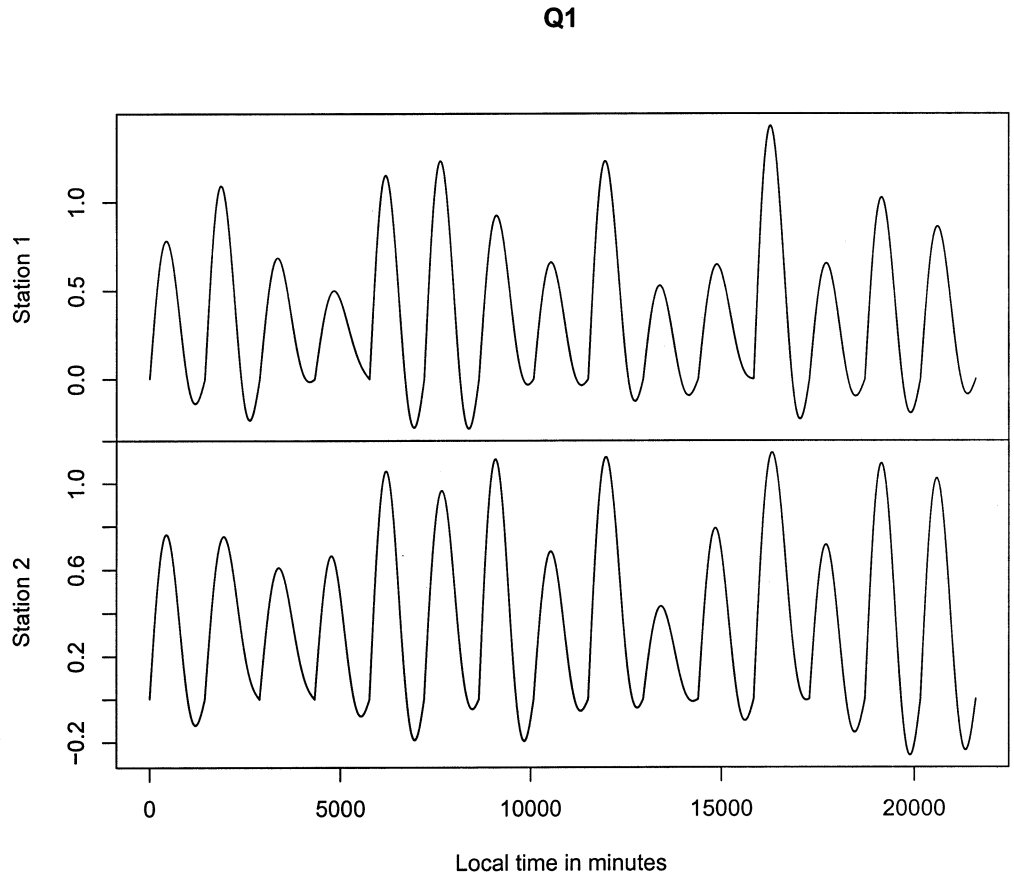


Fig. 5.2: Synthetic "good" Sq example. The most pronounced features are LT aligned.

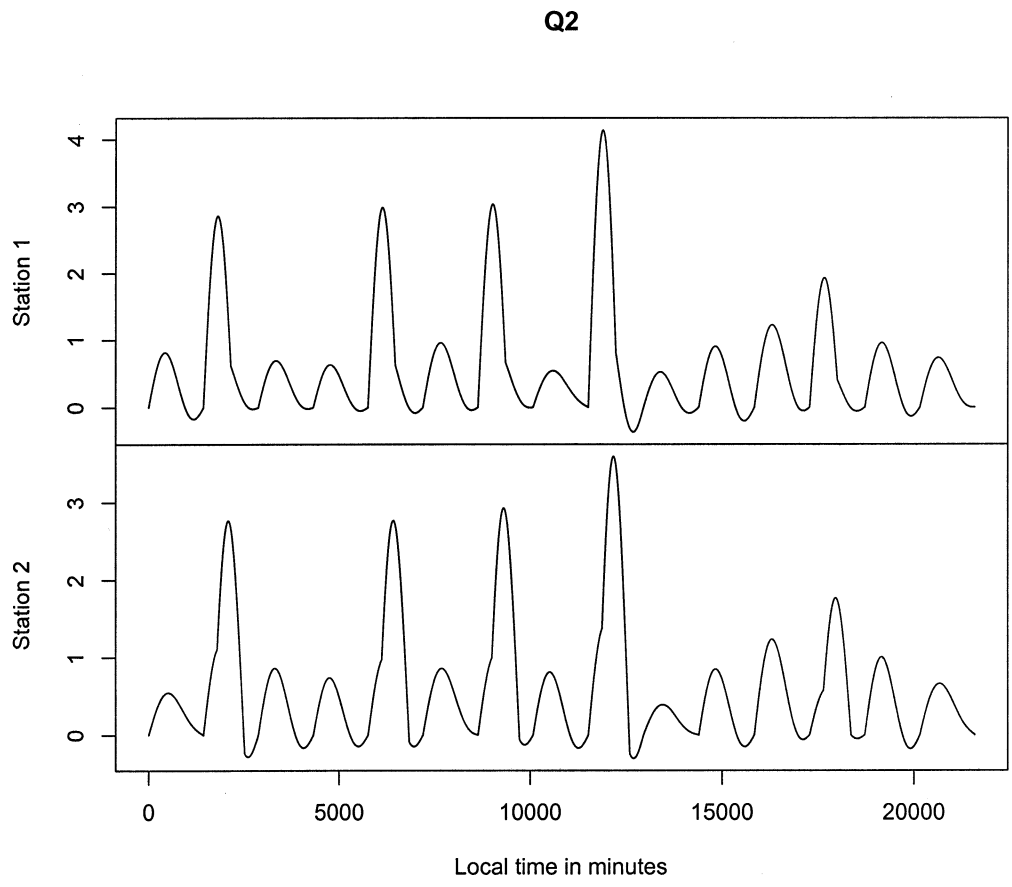


Fig. 5.3: Synthetic "bad" Sq example. Storm features are aligned in UT but shifted in LT.

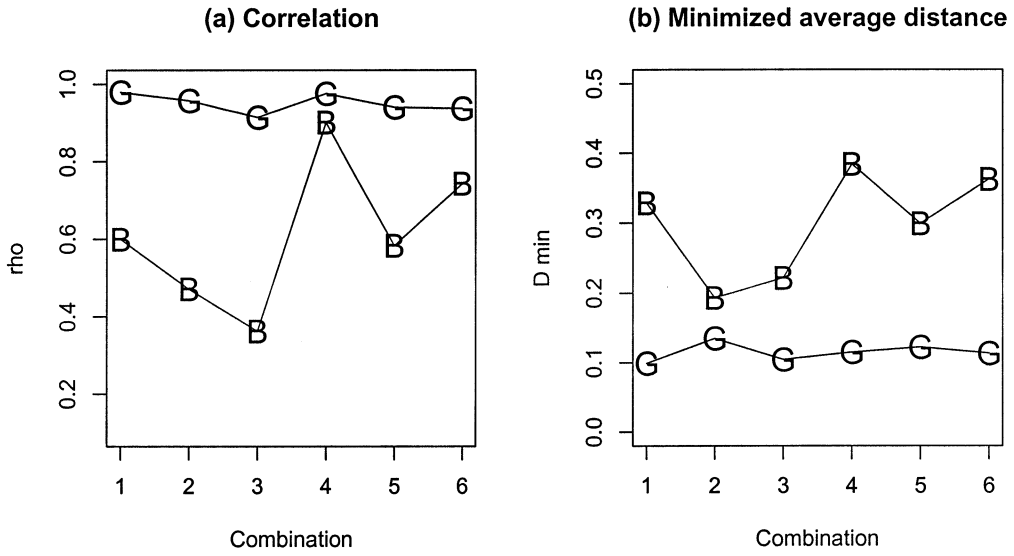


Fig. 5.4: Measures of curve similarity for "good" (G) and "bad" (B) Sq estimates :(a) Correlation, (b) Minimal average distance ($N_D = 15$).

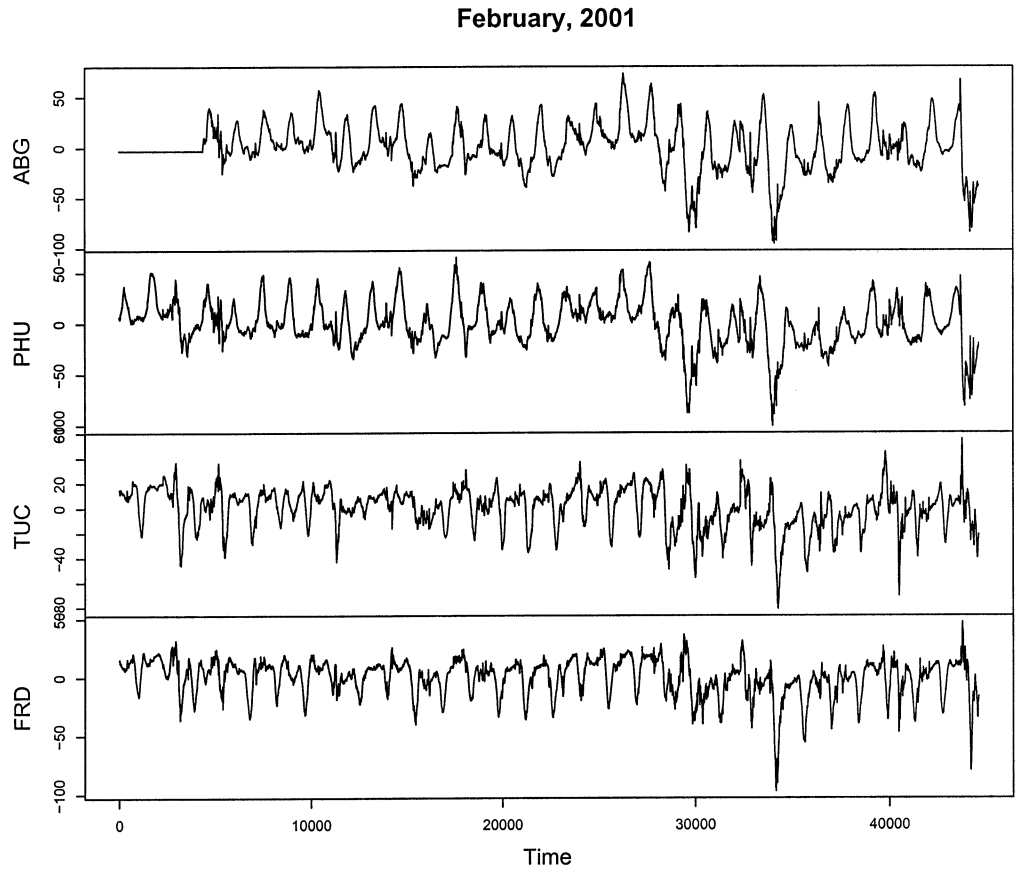


Fig. 5.5: Magnetic field H-component recorded at ABG, PHU, TUC, and FRD stations during February, 2001.

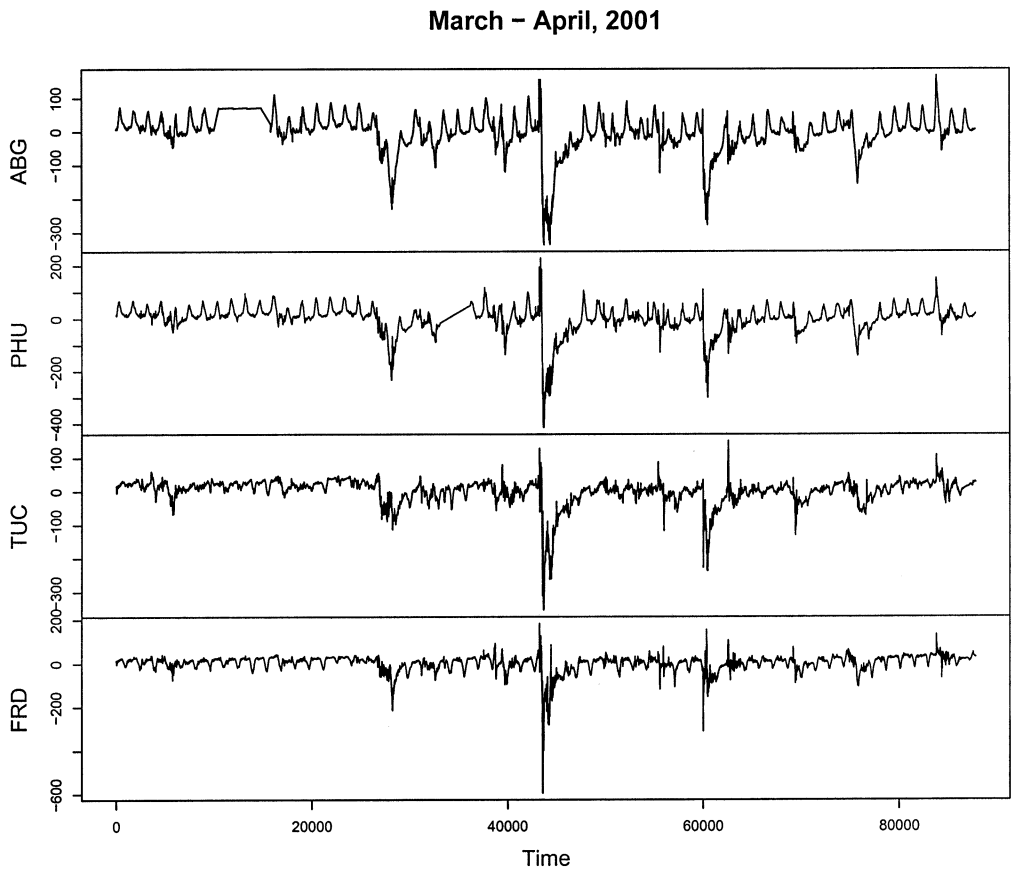


Fig. 5.6: Magnetic field H-component recorded at ABG, PHU, TUC, and FRD stations during March – April, 2001.

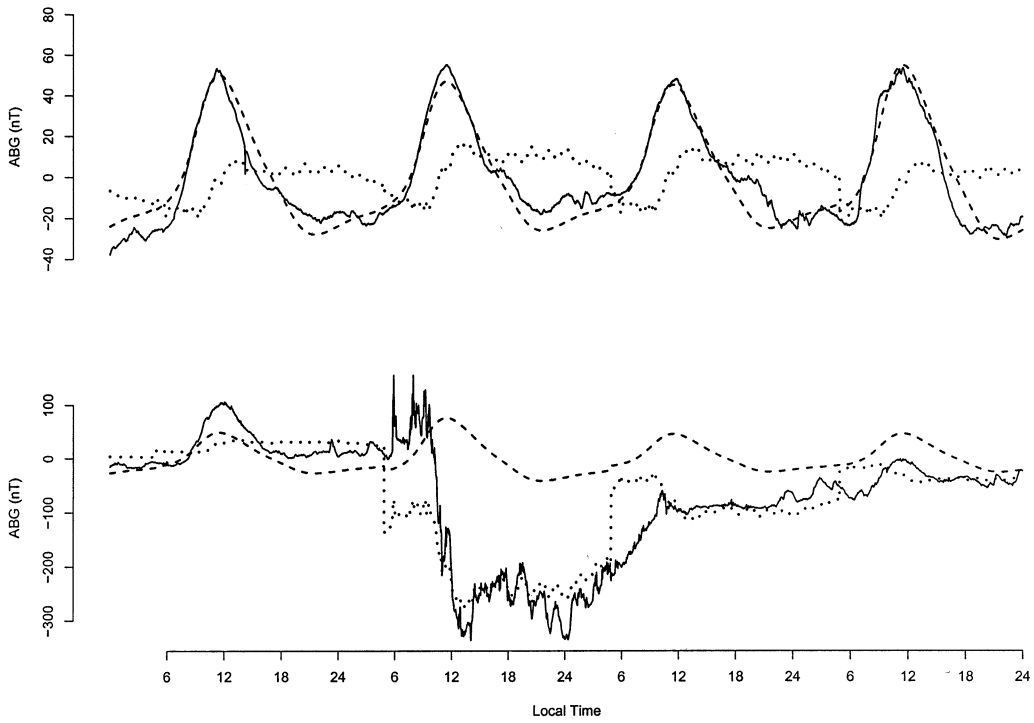


Fig. 5.7: Estimated Sq component using new methodology (dashed line), alternative approach (dotted line), and raw magnetometer data (solid line) at ABG station during quiet period of time: March 14 – March 17, 2001 (top panel) and disturbed period of time: March 29 – April 1, 2001 (bottom panel). The poor performance of the alternative method in the top panel is due to the presence of a storm in the two month period used to construct the estimates. Notice a moderate Sq enhancement of the new Sq estimate that follows the sudden storm commencement (bottom panel)

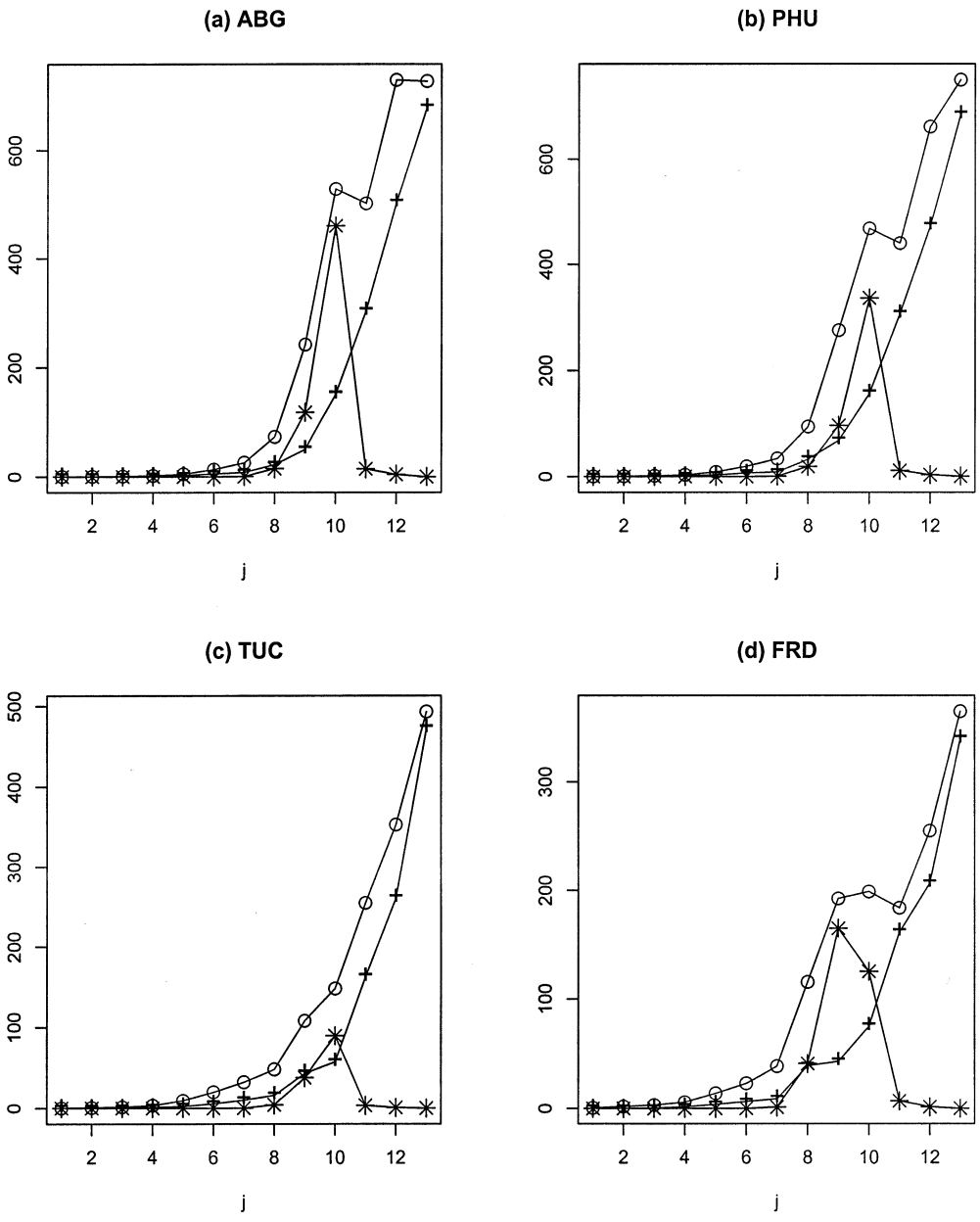


Fig. 5.8: Empirical wavelet power spectra based on MODWT levels $j = 1, \cdots, 13$ for estimated Sq component using new methodology (star), alternative approach (cross), and raw magnetometer data (circle) at (a) ABG, (b) PHU, (c) TUC, and (d) FRD stations during March – April, 2001.

CHAPTER 6

R-PACKAGE

In this chapter we introduce a brief description of the R-package, which is based on the ideas described in Chapters 3 – 5.

The package consists of several functions that are provided in Appendix A and are available at Comprehensive R Archive Network (CRAN). The main functions are `SAIndex` and `SQ`. The first one computes the storm activity index associated with the ring current. It is an automated procedure that requires the user to input the raw magnetometer data and the coordinates of the stations used. The second major function of this package, `SQ`, computes the estimate of the Solar quiet daily variation described in Chapter 5. Further, we provide the details on the use of these functions.

First, we describe basic requirements for function `SAIndex`. In order to compute a storm index one should input the raw magnetometer records and the coordinates of the stations used. One can use any number of roughly equispaced equatorial stations. The input data should be provide as a matrix. Each column of this data matrix contains the records for each individual station. The coordinates should be written in form of a matrix as well. Each column corresponds to the station, the first row must contain the latitude and the second row – longitude. This function returns the global storm index estimate.

In order to estimate the daily nonconstant variations `SQ` function can be used. First, the storm index must be calculated using the function described above. the resulting index is one of the `SQ` inputs. We recommend to use the data from at least two stations to estimate the daily variation. The data should be provided in the matrix form. As mentioned, each column must contain records from different

stations. This functions returns Sq estimated for each station individually.

The functions described in this chapter are easy to use. The only requirement is that the data are provided in the matrix form.

CHAPTER 7

SUMMARY AND CONCLUSIONS

Functional data analysis has many appealing techniques and properties that have great potential in the applied statistics.

In Chapter 2 the test for the lack of dependence for fully functional model is presented. We showed that when the response variable is a functional object the test statistic has χ^2 limiting distribution. The test provides good results (empirical size of the test and power) for samples around 50.

The goal of Chapter 3 is to apply the test for independence introduced in Chapter 2 to magnetometer data. A detailed analysis of the association between auroral currents and the currents at lower latitudes is performed. It indicates that there is a significant association between the substorms recorded at high latitudes and the magnetometer records at mid- and low-latitudes which lasts for 48 hours. The results discovered in this Chapter might imply some physical connections between the substorm electrodynamics and the physical processes in other regions of the Magnetosphere-Ionosphere system that we are not aware of at the present time.

In Chapter 4 an improved procedure of removing the solar quiet daily variations is introduced. We use the wavelet based filtering techniques and the functional principal component analysis to eliminate the nonconstant daily variability and preserve the magnetic storm related features. We developed the procedure that constructs the index that is cleaner than the WISA and the Dst both of which contain significant residual daily variation.

Chapter 5 provides a novel procedure that uses the storm index developed in Chapter 4, multiresolution analysis and functional principal component analysis tech-

niques. We propose the methodology that gives Sq estimate that is stable and extracts daily variations that are not affected by the storm events.

In Chapter 6 a brief description of the R package is provided. Two major functions are developed. One of them, `SAIndex`, computes the storm activity index associated with the ring current. Another function, `SQ`, extracts the Solar quiet (Sq) daily variation from the raw magnetometer data. It uses the storm index `SAIndex` as one of its steps. The main motivation for preparing this package was to make it easier to implement the ideas described in this work. I hope it is going to be a useful tool for space physicists. So, as a part of the future work, I plan to make it more user friendly and interactive.

REFERENCES

[1] B. Liu and H.-G. Müller, "Functional data analysis for sparse auction data," in *Statistical Methods for E-commerce Research*, W. Jank and G. Shmueli, Eds. Wiley, New York, 2008.

[2] M. G. Kivelson and C. T. Russell, Eds., *Introduction to Space Physics.* Cambridge University Press, 1997.

[3] J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis.* Springer Verlag, 2002.

[4] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis.* Springer Verlag, 2005.

[5] J.-M. Chiou, H.-G. Müller, and J.-L. Wang, "Functional response models," *Statistica Sinica*, 14 (2004), pp. 675–693.

[6] N. Malfait and J. O. Ramsay, "The historical functional model," *Canadian Journal of Statistics*, 31 (2003), pp. 115–128.

[7] H. Cardot, F. Ferraty, A. Mas, and P. Sarda, "Testing hypothesis in the functional linear model," *Scandinavian Journal of Statistics*, 30 (2003), pp. 241–255.

[8] H. Cardot, R. Faivre, and M. Goulard, "Functional approaches fro predicting land use with the temproal evolution of coarse resolution remote sensing data," *Journal of Applied Statistics*, 30 (2003), pp. 1185–1199.

[9] H.-G. MÜLLER AND U. STADTMÜLLER, "Generalized functional linear models," *The Annals of Statistics*, 33 (2005), pp. 774–805.

[10] F. YAO, H.-G. MÜLLER, AND J.-L. WANG., "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, 33 (2005), pp. 2873–2903.

[11] T. CAI AND P. HALL, "Prediction in functional linear regression," *The Annals of Statistics*, 34 (2006), pp. 2159–2179.

[12] J.-M. CHIOU AND H.-G. MÜLLER, "Diagnostics for functional regression via residual processes," *Computational Statistics and Data Analysis*, 15 (2007), pp. 4849–4863.

[13] Y. LI AND T. HSING, "On rates of convergence in functional linear regression," *Journal of Multivariate Analysis*, 98 (2007), pp. 1782–1804.

[14] A. CUEVAS, M. FEBRERO, AND R. FRAIMAN, "Linear functional regression: the case of fixed design and functional response," *The Canadian Journal of Statistics*, 30 (2002), pp. 285–300.

[15] H. CARDOT, F. FERRATY, AND P. SARDA, "Spline estimators for the functional linear model," *Statistica Sinica*, 13 (2003), pp. 571–591.

[16] P. KOKOSZKA, I. MASLOVA, J. SOJKA, AND L. ZHU, "Testing for lack of dependence in functional linear model," *Canadian Journal of Statistics*, 36/2 (2008), pp. 207–222.

[17] G. ROSTOKER, "Effects of substorms on the stormtime ring current index Dst," *Annales Geophysicae*, 18 (2000), pp. 1390–1398.

[18] W.-Y. Xu and Y. Kamide, "Decomposition of daily geomagnetic variations by using method of natural orthogonal component," *Journal of Geophysical Research*, 109 (2004), p. A05218, dOI:10.1029/203JA010216.

[19] A. Jach, P. Kokoszka, J. Sojka, and L. Zhu, "Wavelet–based index of magnetic storm activity," *Journal of Geophysical Research*, 111 (2006), p. A09215.

[20] I. Maslova, P. Kokoszka, J. Sojka, and L. Zhu, "Removal of nonconstant daily variation by means of wavelet and functional data analysis," *Journal of Geophysical Research*, 114 (2009), p. A03202, doi:10.1029/2008JA013685.

[21] P. Hall and M. Hosseini-Nasab, "On properties of functional principal components," *Journal of Royal Statistical Society B*, 68 (2006), pp. 109–126.

[22] P. Hall and M. Hosseini-Nasab, "Theory for high–order bounds in functional principal components analysis," The University of Melbourne, Tech. Rep., 2007.

[23] D. Bosq, *Linear Processes in Function Spaces.* New York: Springer, 2000.

[24] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis.* New York: Wiley, 2003.

[25] Y. Kamide, W. Baumjohann, I. A. Danglis, W. D. Gonzalez, M. Grande, J. A. Joselyn, R. L. McPherron, J. L. Phillips, E. G. D. Reeves, G. Rostoker, A. S. Sharma, H. J. Singer, B. T. Tsurutani, and V. M. Vasyliunas, "Current understanding of magnetic storms: Storm–substorm relationships," *Journal of Geophysical Research*, 103 (1998), pp. 17 705–17 728.

[26] I. A. Danglis, J. U. Kozyra, Y. Kamide, D. Vassiliadis, A. S. Sharma, M. Liemohn, W. D. Gonzalez, B. T. Tsurutani, and G. Lu, "Intense space storms: Critical issues and open disputes," *Journal of Geophysical Research*, 108 (2003), p. doi:10.1029/2002JA009722.

[27] I. Maslova, P. Kokoszka, J. Sojka, and L. Zhu, "Effect of substroms on mid- and low-latitude horizontal intensity," Utah State University, Tech. Rep., 2007.

[28] F. Ferraty and P. Vieu, *Nonparametric Functional Data Analysis: Theory and Practice.* New York: Springer Verlag, 2006.

[29] R. Gabrys and P. Kokoszka, "Portmanteau test of independence for functional observations," *Journal of the American Statistical Association*, 102 (2007), pp. 1338-1348, doi:10.1198/016214507000001111.

[30] R. B. Cattell, "The scree test for the number of factors," *Journal of Multivariate Behavioral Research*, 1 (1966), pp. 245–276.

[31] N. A. Tsyganenko, "Modeling the inner magnetosphere: The asymmetric ring current and region 2 Birkland current revisited," *Journal of Geophysical Research*, 105 (2000), p. 27739.

[32] E. Friedrich, G. Rostoker, M. G. Connors, and R. L. McPherron, "Influence of the substorm current wedge on the Dst index," *Journal of Geophysical Research*, 104 (1999), p. 4567.

[33] T. Kikuchi, K. K. Hashimoto, and K. Nozaki, "Penetration of magnetospheric electric fields to the equator during a geomagnetic storm," *Journal of Geophysical Research*, 113 (2008), p. A06214, doi:10.1029/2007JA012628.

[34] K. KITAMURA, H. KAWANO, S. OHTANI, A. YOSHIKAWA, AND K. YUMOTO, "Local time distribution of low and middle latitude ground magnetic disturbances at sawtooth injections of 18-19 april 2002," *Journal of Geophysical Research*, 110 (2005), p. A07208, doi:10.1029/2004JA010734.

[35] P. KOKOSZKA, I. MASLOVA, J. SOJKA, AND L. ZHU, "Testing for lack of dependence in functional linear model," *Canadian Journal of Statistics*, 36 (2008), pp. 207–222.

[36] W. H. CAMPBELL, "Geomagnetic storms: the Dst–ring current myth and log-normal distributions," *Journal of Atmospheric and Terrestial Physics*, 58 (1996), pp. 1171–1187.

[37] G.-X. CHEN, W.-Y. XU, A.-M. DU, Y.-Y. WU, B. CHEN, AND X.-C. LIU, "Statistical characteristics of the day-to day variability in the geomagnetic sq field," *Journal of Geophysical Research*, 112 (2007), p. doi:10.1029/2006JA012059.

[38] J. O. RAMSAY, H. WICKHAM, AND S. GRAVES, *fda: Functional Data Analysis*, 2007, r package version 1.2.3 [Online]. Available: `\tthttp://www.functionaldata.org`.

[39] M. BLANC AND A. D. RICHMOND, "The ionospheric disturbance dynamo," *Journal of Geophysical Research*, 85 (1980), pp. 1669 – 1686.

[40] Z. XU, L. ZHU, J. SOJKA, P. KOKOSZKA, AND A. JACH, "An assessment study of the wavelet-based index of magnetic storm activity (wisa) and its comparison to the dst index," Utah State University, Tech. Rep., 2006.

[41] L. Zhu, Z. Xu, J. Sojka, R. Schunk, P. Kokoszka, and A. Jach, "Are the dst stations sufficient for describing storm-time enhancements," Utah State University, Tech. Rep., 2007.

[42] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis.* Cambridge: Cambridge University Press, 2000.

[43] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* New York: Wiley, 1984.

[44] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, "Canonical correlation analysis when the data are curves," *Journal of the Royal Statistical Society*, 55 (1993), pp. 725–740.

[45] M. Sugiura, "Hourly values of equatorial Dst for the IGY," *Ann. Int. Geophysical Year*, 35 (1964), no. 9, pergamon Press, Oxford.

[46] A. S. Janzhura and O. A. Troshichev, "Determination of the running quiet geomagnetic variation," *Journal of Atmospheric and Solar–Terrestial Physics*, 70 (2008), pp. 962–972.

[47] S. K. Bhardwaj and G. K. Rangarajan, "A model for solar quiet variation at low latitude from past observations using singular spectrum analysis," *Proceedings of the Indian Academy of Sciences (Earth and Planetary Science)*, 107 (1998), pp. 217–224.

[48] E. C. Butcher, "Abnormal Sq behavior," *Pure and Applied Geophysics*, 131 (1989), pp. 463–483.

[49] D. Freedman, R. Pisani, and R. Purves, *Statistics.* New York: W.W. Norton and Company, 1991.

[50] V. P. Golovkov, N. Y. Papitashvili, Y. S. Tyupkin, and Y. P. Kharin, "Separation of geomagnetic field varations into quiet and disturbed components by the method of natural orthogonal components," *Geomag. Aero.*, 18 (1978), pp. 342–344.

APPENDICES

APPENDIX A

R-PACKAGE WAMI CODE

```
SAIndex< −function(data, coord, wf="la8"){
J0< −7
J1< −10
boundary< −"reflection"
periods< −60*c(8,12,24)
SAI< −iWISA(data, wf = "la8", n.levels = n.levels, J0 = J0, J1 = J1, boundary =
boundary, quantile = quantile)
n.station< −dim(data)[2]
N< −dim(data)[1]
index< −matrix(0, ncol=n.station, nrow=N) final.index< −matrix(0, ncol=n.station,
nrow=N)
for(i in 1:n.station){
index[,i]< −SAI$preindex[,i]/cos(magnetic.latitude(coord[1,i],coord[2,i])$Phi*2*pi/360)}
for(i in 1:n.station){final.index[,i]< −index[,i]-mean(index[,i])}
SI< −numeric(N)
SI< −apply(final.index, 1, mean)
return(SI=SI) } }
library(waveslim)

iWISA< −function(data, wf = "la8", n.levels = 11, J0 = 7, J1 = 12, boundary =
"reflection", quantile = 0.9) {
if(is.matrix(data)==F & is.array(data)==F & is.data.frame(data)==F) stop("wrong
```

```
data format: input data should be matrix or array") else {
n.station< −dim(data)[2]
N< −dim(data)[1] if(N>=(365*1440)) n.levels< −floor(log(N,2))
preindex < − matrix(data = 0, ncol = n.station, nrow = N)
mra.sq < − matrix(data = 0, ncol = n.station, nrow = N)
for(i in 1:n.station){
data.wt < − modwt(data[, i], wf=wf, n.levels=n.levels, boundary=boundary)
which.levels< −1:J0
data.wt.th< −quantile.manual.thresh.scalewise(data.wt, which.levels=which.levels, hard=FALSE,
quantile=quantile)
data.wt.th.mra< −mra.wt(data.wt.th)
data.recon< −numeric(N)
for(f in 1:J0) data.recon< −data.recon+data.wt.th.mra[[f]]
for(f in (J0+1):J1) mra.sq[,i] < − mra.sq[,i] + data.wt.th.mra[[f]]
for(j in (J1+1):n.levels)
data.recon< −data.recon+data.wt.th.mra[[j]]
smooth< −data.wt.th.mra[[n.levels+1]]
if(N>=(365*1440)) data.recon< −data.recon+mean(smooth)
if(N¡(365*1440)) data.recon< −data.recon+smooth
preindex[,i]< −preindex[,i]+data.recon }
deco< −rem.daily(mra.sq)
preindex< −preindex+deco$recon
return(preindex = preindex, pseudo.sq = deco$SQ, smooth = smooth)} }

rem.daily< −function(data){
n.station< −dim(data)[2]
```

```
N< −dim(data)[1]
recon< −matrix(0, ncol = n.station, nrow = N)
mean.v< −numeric(N)
if(n.station==1) mean.v< −numeric(N) else mean.v< −apply(data, 1, mean)
data.centr< −matrix(NA, ncol=n.station, nrow = N)
for(i in 1:n.station){data.centr[,i]< −data[,i] - mean.v}
SQ < − deco$sq for(i in 1:n.station) recon[,i] < − deco$diff[,i] + mean.v
return(recon = recon, SQ = SQ)}

pca.sq< −function(data){
period = 1440
n.station< −dim(data)[2]
N< −dim(data)[1]
how.many< −floor(N/period)
z< −matrix(data = NA, ncol = n.station, nrow = (how.many*period))
for(i in 1:n.station){
z[, i]< −data[1:(how.many*period), i]}
z.new< −array(NA, dim=c( period, how.many,n.station))
for(k in 1:n.station){ z.new[, , k]< − matrix(z[ ,k], nrow=period)}
scores< −matrix(NA, ncol=n.station, nrow = how.many)
harm< −matrix(NA, ncol=n.station, nrow = period)
d< −create.bspline.basis(rangeval=c(0, period), nbasis=159)
t< −seq(1, period, 1)
for(k in 1:n.station){
fd< −data2fd(z.new[,,k],t,basisobj=d)
pca< −pca.fd(fd, nharm = 1, centerfns = F)
```

```
scores[, k]< −pca$scores[,1]
harm[, k]< −eval.fd(c(1:period), pca$harmonics[1])}
q< −numeric(n.station)
for(i in 1:n.station) q[i]< −quantile(abs(scores[,i]), probs=0.95)
med< −numeric(n.station)
for(i in 1:n.station) {med[i]< −median(scores[,i])}
delta< −matrix(0, ncol=n.station, nrow = how.many)
for(j in 1:n.station){
for(i in 1:how.many){
if(abs(scores[i,j]) > q[j]) delta[i,j]< −1 }}
ind< −apply(delta,1,sum)
for(j in 1:n.station){
for(i in 1:how.many){
if(ind[i]==(n.station)) scores[i,j]< −med[j] }}
SQ< −matrix(data = NA, ncol = n.station, nrow = N)
for(j in 1:n.station){
for(i in 1:how.many){
s< −period*(i-1)+1
e< −period*i
d.sq< −(harm[,j]*scores[i,j])
SQ[s:e,j]< −d.sq } }
list(sq=SQ, diff=(data-SQ)) }

quantile.manual.thresh.scalewise< −function(wc, which.levels=c(1,2,3), hard = T,
quantile=.9) { wc.shrink < − wc
if (hard) { for (i in names(wc)[which.levels]) {
```

```
wci < − wc[[i]]
unithresh < − quantile(abs(wci),quantile)
wc.shrink[[i]] < − wci *(abs(wci) > unithresh) } }
else { for (i in names(wc)[which.levels]) {
wci < − wc[[i]]
unithresh < − quantile(abs(wci),quantile)
wc.shrink[[i]] < − sign(wci)* (abs(wci) - unithresh) * (abs(wci) > unithresh) } }
wc.shrink }

mra.wt< −function(x.wt) {
wf< −attr(x.wt,"wavelet")
J< −length(x.wt)-1
method< −attr(x.wt,"class")
boundary< −attr(x.wt,"boundary")
if(method=="modwt") n< −length(x.wt[[1]]) else n< −2*length(x.wt[[1]])
x.mra < − vector("list", J + 1)
zero < − vector("list", J + 1)
names(zero) < − c(paste("d", 1:J, sep = ""), paste("s", J, sep = ""))
class(zero) < − method
attr(zero, "wavelet") < − wf
attr(zero, "boundary") < − boundary
zero[[J + 1]] < − x.wt[[J + 1]]
if (method == "modwt") { for (k in 1:J) zero[[k]] < − numeric(n)
x.mra[[J + 1]] < − imodwt(zero) }
else { for (k in 1:J) zero[[k]] < − numeric(n/2k̂)
x.mra[[J + 1]] < − idwt(zero) }
```

```
for (j in J:1) { zero < − vector("list", j + 1)
names(zero) < − c(paste("d", 1:j, sep = ""), paste("s", j, sep = ""))
class(zero) < − method
attr(zero, "wavelet") < − wf
attr(zero, "boundary") < − boundary
zero[[j]] < − x.wt[[j]]
if (method == "modwt") { if (j != 1) {
for (k in c(j + 1, (j - 1):1)) zero[[k]] < − numeric(n) }
else { zero[[j + 1]] < − numeric(n) }
x.mra[[j]] < − imodwt(zero) }
else { zero[[j + 1]] < − numeric(n/2ĵ)
if (j != 1) { for (k in (j - 1):1) zero[[k]] < − numeric(n/2k̂)}
x.mra[[j]] < − idwt(zero) } }
names(x.mra) < − c(paste("D", 1:J, sep = ""), paste("S", J, sep = ""))
if (boundary == "reflection") { for (j in (J + 1):1) x.mra[[j]] < − x.mra[[j]][1:(n/2)]
return(x.mra) }
else { return(x.mra) } }
    SQ< −function(data, si.v=si.v, wf = "la8", n.levels = 10, boundary ="reflec-
tion"){
J0< −7
J1< −10
if(is.vector(data)==T) {n.station< −1 N< −length(data)} else {n.station< −dim(data)[2]
N< −dim(data)[1]
library(waveslim)
data.ds< −matrix(NA, ncol=n.station, nrow = N)
for(i in 1:n.station){data.ds[,i]< −data[,i] - si.v}
```

```
mra.sq < − matrix(data = 0, ncol = n.station, nrow = N)
for(i in 1:n.station){ data.mra< −mra(data.ds[,i], wf=wf, J = n.levels, boundary=boundary)
for(f in (J0+1):J1) mra.sq[,i] < − mra.sq[,i] + data.mra[[f]] }
deco< −pca.SQ.new(mra.sq)
SQ < − deco$sq return(SQ) }
    pca.SQ.new< −function(data) {
period = 1440
n.station< −dim(data)[2]
N< −dim(data)[1]
how.many< −floor(N/period)
z< −matrix(data = NA, ncol = n.station, nrow = (how.many*period))
for(i in 1:n.station){z[, i]< −data[1:(how.many*period), i]}
z.new< −array(NA, dim=c( period, how.many,n.station))
for(k in 1:n.station){z.new[, , k]< − matrix(z[ ,k], nrow=period)}
scores< −matrix(NA, ncol=n.station, nrow = how.many)
harm< −matrix(NA, ncol=n.station, nrow = period)
d< −create.bspline.basis(rangeval=c(0, period), nbasis=159)
t< −seq(1, period, 1)
for(k in 1:n.station){
fd< −data2fd(z.new[,,k],t,basisobj=d)
pca< −pca.fd(fd, nharm = 1, centerfns = F)
scores[, k]< −pca$scores[,1]
harm[, k]< −eval.fd(c(1:period), pca$harmonics[1])}
q< −numeric(n.station)
for(i in 1:n.station) q[i]< −quantile(abs(scores[,i]-median(scores[,i])), probs=0.9)
med< −numeric(n.station)
```

```
for(i in 1:n.station) med[i]< −median(scores[,i])
delta< −matrix(0, ncol=n.station, nrow = how.many)
for(j in 1:n.station){ for(i in 1:how.many){
if(abs(scores[i,j]-median(scores[,j])) ¿ q[j]) delta[i,j]< −1 }}
ind< −apply(delta,1,sum)
for(j in 1:n.station){ for(i in 1:how.many){
if(ind[i]==(n.station)) scores[i,j]< −med[j]}}
SQ< −matrix(data = NA, ncol = n.station, nrow = N)
for(j in 1:n.station){ for(i in 1:how.many){
s< −period*(i-1)+1
e< −period*i
d.sq< −(harm[,j]*scores[i,j])
SQ[s:e,j]< −d.sq} }
list(sq=SQ, diff=(data-SQ)) }
```

APPENDIX B

PERMISSIONS

April 17th 2009

Inga Maslova
Department of Mathematics and Statistics
3900 Old Main Hill
Logan, UT 84322-3900
Phone: (435)-757-0548

Dear Jan Sojka:

I am in the process of preparing my dissertation in the Department of Mathematics and Statistics at Utah State University. I hope to complete it in the summer of 2009. A am requesting your permission to include the attached material as shown. I will include acknowledgments and/or appropriate citations to your work as shown and copyright and reprint rights information in a special appendix. The bibliographical citation will appear at the end of the manuscript as shown. Please advise me of any changes you require.

Please indicate your approval of this request by signing in the space provided, attaching any other form or instruction necessary to confirm permission. If you charge a reprint fee for use of your material, please indicate this as well. If you have any questions, please call me at the number above.

I hope you will be able to reply immediately. If you are not the copyright holder, please forward my request to the appropriate person or institution.

Thank you for your cooperation,

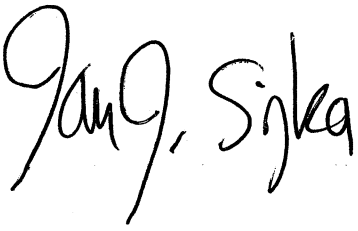
Inga Maslova

I hereby give permission to Inga Maslova to reprint the following material in her dissertation.

Chapter 2 of this dissertation based on paper by P. Kokoszka, I. Maslova, J. Sojka, L. Zhu, Testing for lack of dependence in the functional linear model, Canadian Journal of Statistics, Vol. 36, No 2, 2008.

Chapter 3 of this dissertation based on manuscript by I. Maslova. P. Kokoszka, J. Sojka, and L.Zhu, Statistical significance testing for the association of magnetometer records at high–, mid– and low latitudes during substorm days.

Chapter 4 of this dissertation based on paper by I. Maslova, P. Kokoszka, J. Sojka, and L. Zhu, Removal of nonconstant daily variation by means of wavelet and functional data analysis, Journal of Geophysical Research, Vol. 114, doi:10.1029/2008JA013685, 2009.

Signed 

June 1st 2009

Inga Maslova
Department of Mathematics and Statistics
3900 Old Main Hill
Logan, UT 84322-3900
E-mail: Inga.Maslova@gmail.com
Phone: (435)-757-0548
Fax: (435)-797-1822

Canadian Journal of Statistics
Phone (604-822-1300)
Fax (604-822-6960)
E - mail: cjs@stat.ubc.ca

To Paul Gustafson:

I am preparing my dissertation in the Department of Mathematics and Statistics at Utah State University. I hope to complete degree in the Summer of 2009.

An article, Testing for Lack of Dependence in the Functional Linear Model, of which I am the second author, and which appeared in your journal Vol. 36, No 2, pages 207–222, 2008, reports an essential part of my dissertation research. I would like permission to reprint it as a chapter in my dissertation. (Reprinting the chapter may necessitate some revision.) Please note that USU sends dissertations to Bell & Howell Dissertation Services to be made available for reproduction.

I will include an acknowledgment to the article on the first page of the chapter, as shown below. Copyright and permission information will be included in a special appendix. If you would like a different acknowledgment, please so indicate.

Please indicate your approval of this request by signing in the space provided, and attach any other form necessary to confirm permission. If you charge a reprint fee for

use of an article by the author, please indicate that as well.

If you have any questions, please call me at the number above or send me an e-mail message at the above address. Thank you for your assistance.

Inga Maslova

I hereby give permission to Inga Maslova to reprint the requested article in her dissertation, with the following acknowledgment:

(P. Kokoszka, I. Maslova, J. Sojka, L. Zhu, Testing for Lack of Dependence in the Functional Linear Model, Canadian Journal of Statistics, Vol. 36, No 2, pages 207–222, 2008)

Signed [signature], Managing Editor the Canadian Journal of Statistics

Date 12 June 2009

June 1st 2009

Inga Maslova
Department of Mathematics and Statistics
3900 Old Main Hill
Logan, UT 84322-3900
E-mail: Inga.Maslova@gmail.com
Phone: (435)-757-0548
Fax: (435)-797-1822

Journal of Geophysical Research - Space Physics
Publications Administration
American Geophysical Union
2000 Florida Avenue, N.W. Washington,
DC 20009

To Zuyin Pu:

I am preparing my dissertation in the Department of Mathematics and Statistics at Utah State University. I hope to complete degree in the Summer of 2009.

An article, Removal of Nonconstant Daily Variation by Means of Wavelet and Functional Data Analysis, of which I am the first author, and which appeared in your journal Vol. 114, A03202, doi:10.1029/2008JA013685, 2009, reports an essential part of my dissertation research. I would like permission to reprint it as a chapter in my dissertation. (Reprinting the chapter may necessitate some revision.) Please note that USU sends dissertations to Bell & Howell Dissertation Services to be made available for reproduction.

I will include an acknowledgment to the article on the first page of the chapter, as shown below. Copyright and permission information will be included in a special appendix. If you would like a different acknowledgment, please so indicate.

Please indicate your approval of this request by signing in the space provided, and attach any other form necessary to confirm permission. If you charge a reprint fee for use of an article by the author, please indicate that as well.

If you have any questions, please call me at the number above or send me an e-mail message at the above address. Thank you for your assistance.

Inga Maslova

I hereby give permission to Inga Maslova to reprint the requested article in her dissertation, with the following acknowledgment:

(I. Maslova, P. Kokoszka, J. Sojka, L. Zhu, Removal of Nonconstant Daily Variation by Means of Wavelet and Functional Data Analysis, Journal of Geophysical Research, Vol. 114, A03202, doi:10.1029/2008JA013685, 2009)

Signed

Date 8 Jan 2009

please see attached letter

We are pleased to grant permission for the use of the material requested for inclusion in your thesis. The following non-exclusive rights are granted to AGU authors:

- All proprietary rights other than copyright (such as patent rights).
- The right to present the material orally.
- The right to reproduce figures, tables, and extracts, appropriately cited.
- The right to make hard paper copies of all or part of the paper for classroom use.
- The right to deny subsequent commercial use of the paper.

Further reproduction or distribution is not permitted beyond that stipulated. The copyright credit line should appear on the first page of the article or book chapter. The following must also be included, Reproduced by permission of American Geophysical Union. To ensure that credit is given to the original source(s) and that authors receive full credit through appropriate citation to their papers, we recommend that the full bibliographic reference be cited in the reference list. The standard credit line for journal articles is: "Author(s), title of work, publication title, volume number, issue number, citation number (or page number(s) prior to 2002), year. Copyright [year] American Geophysical Union."

If an article was placed in the public domain, in which case the words Not subject to U.S. copyright appear on the bottom of the first page or screen of the article, please substitute published for the word copyright in the credit line mentioned above.

Copyright information is provided on the inside cover of our journals. For permission for any other use, please contact the AGU Publications Office at AGU, 2000 Florida Ave., N.W., Washington, DC 20009.

Michael Connolly

Journals Publications Specialist

Thank you for your interest in reproducing AGU published material. AGU does not require that permission be obtained from AGU or the author(s) for the use of tables, figures, or short extracts of papers published in AGU journals or books, provided that the original publication be appropriately cited.

The standard credit line for the citation is, Author(s), title, publication, volume number, issue number, citation number (or page number(s) prior to 2002), date. Copyright [year] American Geophysical Union. The following must also be included, Reproduced/modified by permission of American Geophysical Union.

If an article was placed in the public domain, in which case the words Not subject to U.S. copyright appear on the bottom of the first page or screen of the article, please substitute published for the word copyright in the credit line mentioned above.

Copyright information is provided on the inside cover of our journals. For permission for any other use, please contact the AGU Publications Office at AGU, 2000 Florida Ave., N.W., Washington, DC 20009.

Michael Connolly

Journals Publications Specialist

CURRICULUM VITAE

Inga Maslova

EDUCATION:

PhD (Statistical Science), May 2009, Utah State University, USA (Professor P. Kokoszka)
M.S., (Statistics), May 2005, Utah State University, USA (Professor P. Kokoszka)
B.S. (Statistics), May 2003, Cum laude diploma, Vilnius University, Lithuania (Professor B. Grigelionis)

PROFESSIONAL EXPERIENCE:

Graduate Research Assistant, WAMI (Wavelet Analysis of Magnetosphere-Ionosphere) Project, 2006-2009, Utah State University
Graduate Teaching Instructor, 2003-2009, Utah State University

PUBLICATIONS:

I. Maslova. P. Kokoszka, J. Sojka, and L.Zhu, Effect of substorms on the magnetic field variability at mid- and low-latitudes, *under review*, 2009

I. Maslova, P. Kokoszka, J. Sojka, and L. Zhu, Removal of nonconstant daily variation by means of wavelet and functional data analysis, *Journal of Geophysical Research*, 114, A03202, doi:10.1029/2008JA013685, 2009.

P. Kokoszka, I. Maslova, J. Sojka, L. Zhu, Testing for lack of dependence in the functional linear model, Canadian Journal of Statistics, Vol. 36, No 2, 207 - 222, 2008

P. Kokoszka, I. Maslova, J. Sojka, L. Zhu, Probability tails of wavelet coefficients of magnetometer records, Journal of Geophysical Research-Space Physics, Vol. 111, No. A6, A06202, 10.1029/2005JA011486, 2006

J. Sojka, A. Jach, P. Kokoszka, I. Maslova, L. Zhu, Z. Xu, Statistical wavelet analysis of magnetometer data: probability tails and geomagnetic storm index, forthcoming

Baker, M. Jung, Ch. Lee, I. Maslova, M. Morton, J. Wang, Analysis of biological interaction networks for drug discovery, CRSC Technical Report (CRSC-TR06-23), 2006